Designing Artificial Identity: The Identity Design Framework and Research Agenda

Karla Bransky
karla.kelly@anu.edu.au
The Australian National University
Canberra, Australia

Penny Sweetser
penny.kyburz@anu.edu.au
The Australian National University
Canberra, Australia

Patrick Holthaus
p.holthaus@herts.ac.uk
University of Hertfordshire
Hatfield, United Kingdom

Guy Laban
laban@bgu.ac.il
Ben-Gurion University of the Negev
Beer Sheva, Israel

Katie Winkle
katie.winkle@it.uu.se
Uppsala University
Uppsala, Sweden

Neziha Akalin
neziha.akalin@ju.se
Jönköping University
Jönköping, Sweden

Ashita Ashok
ashita.ashok@cs.rptu.de
University of Kaiserslautern-Landau
Kaiserslautern, Germany

Jihye Lee
jihyelee.123@snu.ac.kr
Seoul National University
Seoul, Republic of Korea

Rucha Khot
r.khot@tue.nl
Eindhoven University of Technology
Eindhoven, Netherlands

Alexandra Bejarano
abejarano@vt.edu
Virginia Tech
Blacksburg, United States of America

Jorrit Thijn
jorrit.thijn@hku.nl
University of the Arts
Utrecht, Netherlands

Roger K. Moore
r.k.moore@sheffield.ac.uk
University of Sheffield
Sheffield, United Kingdom

Minsu Jang
minsu@etri.re.kr
Electronics and Telecommunications Research Institute
Daejeon, South Korea

Joel Fischer
joel.fischer@nottingham.ac.uk
University of Nottingham
Nottingham, United Kingdom

Minha Lee
m.lee@tue.nl
Eindhoven University of Technology
Eindhoven, Netherlands

Abstract

The identity design of artificial agents carries growing ethical, psychological, and cultural weight, as ubiquitous language models and diverse robotic forms are blended into everyday use. However, structured approaches to designing coherent and interpretable artificial identities remain limited. To address urgent challenges in artificial identity design, including harmful stereotypes and deceptive practices, we introduce the Identity Design (ID) Framework and an accompanying research agenda. Drawing on emerging work on artificial identity in human-robot interaction and taking an interdisciplinary perspective, we propose twelve design principles across three levels: individual (recognisability, behavioural consistency, identity continuity, memory, persistent goals), group (membership signalling, social alignment, role clarity), and societal (benevolence, artificiality, social justice, transparency). The research agenda outlines open questions around the operationalisation and measurement of identity, social dynamics, and ethical considerations for identity design. Together, they lay the groundwork for future research and responsible practice in robotic, virtual, and multi-embodied agents.

CCS Concepts

• **Human-centered computing → HCI theory, concepts and models**.

Keywords

Artificial Identity, Artificial Agents, Social Robots, Identity Design, Design Principles, Multi-Embodiment, Entitativity

ACM Reference Format:
Karla Bransky, Penny Sweetser, Patrick Holthaus, Guy Laban, Katie Winkle, Neziha Akalin, Ashita Ashok, Jihye Lee, Rucha Khot, Alexandra Bejarano, Jorrit Thijn, Roger K. Moore, Minsu Jang, Joel Fischer, and Minha Lee. 2026. Designing Artificial Identity: The Identity Design Framework and Research Agenda. In *Proceedings of the 2026 ACM International Conference on Designing Interactive Systems (DIS '26).* ACM, New York, NY, USA, 18 pages. https://doi.org/XXXXXXX.XXXXXXX

DIS '26, Singapore

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Identity design is recognised as a critical concern in the development of social robots and artificial agents [32, 102, 116, 123]. Recent advances in artificial intelligence (AI), including generative AI and large language models (LLMs) have democratised access

to social agent creation, enabling users to interact with and personalise artificial agents through platforms such as ChatGPT, Replika, and Character.AI [53, 80]. While physically embodied social robots remain less commercially widespread than conversational agents, the deployment of adaptive identity via AI technologies in robotic agents is growing due to their social and affective potential [139, 193, 209]. Robotic platforms like Furhat now support rapid customisation of embodied agent identities, and LLMs are increasingly being integrated into robotic systems such as Ameca [13], Pepper [83], QTrobot [112, 118], or Spot [137], to support open-ended interaction. Together, these developments mark a shift in how artificial identity is designed, enacted, and experienced.

This shift raises new design challenges, especially as designerly human-computer interaction (HCI) and human-robot interaction (HRI) research stands to be expanded [134]. The construction of artificial identity was once the exclusive domain of system designers and engineers, but everyday people now participate in shaping agent personas. People value the ability to customise features such as name, voice, and personality of artificial agents [80, 172]. However, as these technologies become more sophisticated and socially embedded in multi-user and multi-agent contexts, the identity cues they project and signals they emit, across virtual and embodied forms, carry increasing ethical, psychological, and cultural weight [113, 141, 215, 222]. In particular, the rise of multi-embodied systems, in which a single agent may transition between multiple appearances, platforms, or roles, demands more robust frameworks for designing and evaluating identity continuity and coherence of social agents [32, 119, 123].

In this work, as an interdisciplinary cohort of HCI and HRI researchers across engineering, design, philosophy, psychology, arts, computer science, and social sciences, we present the Identity Design (ID) framework for artificial agents. The ID framework consists of a layered set of 12 design principles [76] and guiding questions to support the intelligible and responsible design of artificial identity for social robots and artificial agents. We propose these principles as preliminary, evidence-informed guidance for artificial identity design, across three design lenses, designing for interaction at the individual, group, and societal level.

Drawing on research in robo-identity and social psychology, including social identity and entitativity ("groupness"), the framework provides a conceptual foundation for creating recognisable, coherent, socially meaningful, and ethically informed artificial identities. It is intended as a theoretical framework for researchers and designers of robotic, virtual, or multi-embodied agents, provoking reflection and guiding future work in identity design. In addition, we outline a research agenda to advance systematic study and responsible practice in this emerging research area.

2 Related Work

The ID framework builds on the core concepts of artificial agents, artificial identity, and multi-embodiment. In this section, we first outline these concepts and then discuss the need for identity design principles and the motivation for the framework.

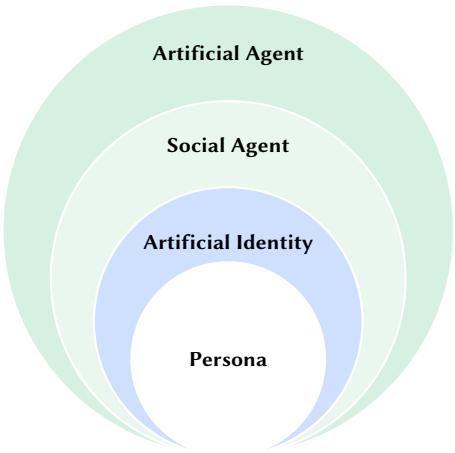

Figure 1: Conceptual Scope of Artificial Agent, Social Agent, Artificial Identity, and Persona

2.1 Artificial Agents

An artificial agent is a non-human, technological system that can act as an interaction partner. This includes machines, robots, chatbots, conversational agents, virtual agents, and interactive, conversational, and affective AIs (alongside variations of such terms). Thus, artificial agents can be visibly embodied (e.g., physical robots and embodied virtual agents) or they may not rely on a physical form (e.g., chatbots and virtual assistants). Artificial agents may also appear as a part of multi-embodiment systems interacting across forms, environments, and contexts (see Section 2.3). While debates continue in HRI and related fields about distinctions between terms such as bots, chatbots, and conversational agents, we have a broad view, since all of these non-human systems can serve as our interaction partners.

People often interact socially with artificial agents; the CASA paradigm demonstrates that humans apply social scripts to computers, treating them as conversational partners, teammates, or even moral actors [125, 144, 152, 153]. Psychological research on anthropomorphism further shows that humans readily attribute emotions, intentions, behaviours, and physical characteristics to non-human agents [19, 63, 213]. The term *social agent* highlights these social dimensions [187], and is nested under our usage of *artificial agent* as a broad, inclusive term. It encompasses socially perceived behaviour as well as computational, goal-driven forms of agency defined in relation to human purposes [18, 67]. For instance, agency is observer-dependent according to levels of abstraction, [100], varying with interactivity, autonomy, and adaptability [66, 69]. This perspective accommodates interactions in which machines are perceived as subjects rather than objects, allowing morally relevant engagements from a second-person point of view [48, 124, 196]. Crucially, what matters for understanding artificial identity is not merely embodiment, role, or context, but also how a system presents itself and is perceived—its persona or metaphorical 'skin' [44, 146].

2.2 What is Artificial Identity?

When discussing an artificial agent's 'identity', we are attentive to the distinction between the raw physical artefact, how it portrays itself in its environment, and how it is perceived by observers in that environment [146]. That is, the persona depicted by an artificial agent is conditioned on its intended role, which depends on the given use case, e.g., a robot acting socially to greet people at airports. How convincingly the artificial agent presents itself will affect whether people accept it as a social partner or see it as just a machine [44].

Recently, HRI has begun to formalise artificial identity. Miranda [140] frames identity as an emergent, relational configuration rather than fixed properties of an entity, drawing on Buddhist philosophy. In this model, identity consists of five interdependent aggregates of *form, sensory input, labelling, assemblage,* and *consciousness*, that are spatially and temporally composable. Characteristics like gender, role, or embodiment are positioned not as inherent properties, but as perceived constructs shaped by context and interaction. Seaborn [187] complements this view, adapting the Social Identity Approach (SIA) [201] to artificial agents, arguing that robots participate in identity processes such as social categorisation, identification, and comparison. While robots are typically externally categorised by users, Seaborn suggests that future agents may engage in self-identification by asserting, negotiating, or resisting group memberships. Both approaches foreground the social and ethical dimensions of agent identity, highlighting risks such as identity multiplicity, social favouritism, and representational bias. Miranda and Seaborn offer complementary perspectives for understanding the complexities of identity design.

Historically, in HCI and design, personas are pragmatic design tools [167]; fictional yet evidence-based user archetypes used to capture user goals and needs rather than to specify an agent's identity. More recently, in conversational AI, the term persona often refers to the 'front-facing' design layer, a set of agent characteristics that shape agent style or behaviour. Here, we define artificial identity as *subsuming* persona: it encompasses the intended (by designers and engineers), projected (by users), and perceived (by humans and non-human interactants) aspects of an agent, all of which evolve over time (see Figure 1).

2.3 Multi-Embodiment

Artificial agents are not confined to a single form, but can extend across multiple embodiments, creating new opportunities for adaptable, cross-context interaction and persistent identity. This capacity, termed *multi-embodiment*, allows agents to transition between physical and virtual embodiments or even occupy multiple forms simultaneously [32, 123]. Luria et al. [135] identified four main configurations: *one-for-one*, in which a single identity inhabits a single embodiment; *one-for-all*, where a single identity is perceived to operate across many embodiments (e.g., Alexa or Siri); *re-embodiment*, in which a single agent transitions between embodiments over time; and *co-embodiment*, where multiple agents co-inhabit a single body.

Re-embodiment has received the most sustained attention in HRI [32]. Preserving a coherent and recognisable identity across embodiments is a core challenge in such configurations [108, 138, 172, 191]. Martin et al. [138] introduced the concept of *identity cues*; stable characteristics that aid recognition across embodiments and help users maintain a sense of continuity. Identity cues include features such as form, appearance, name, voice, and behaviours. Laity et al. [119] reviewed identity and migration cues in multi-embodiment literature and proposed a framework for persona design in migrating agents, emphasising intentional strategies to ensure agents remain recognisable and trusted across embodiments.

Co-embodiment also introduces complexity for identity design, as the issue is not just whether agents sharing a body are perceived as separate entities or as a unified whole, but whether, and how, designers can actively shape that perception [43, 172]. For example, when a human and an artificial agent jointly operate the same robot, designers may wish to control whether other users or observers interpret the embodiment as representing one operator or multiple. This raises important questions about which communicative signals or behaviours can make such distinctions clear.

Entitativity or "groupness" is also relevant for identity design and multi-embodiment, referring to the extent to which a group appears to be a unified whole [4, 20, 41, 71]. Just as groups of people or animals vary to the degree they appear cohesive and coherent [131], agents can also be interpreted as being more or less coherent units. This perception has direct implications for identity design; the way robots signal shared purpose, synchrony, or common attributes shapes whether they are seen as functioning collectives or as disconnected individuals [22, 71]. As such, the group identity conveyed through entitativity becomes a central design consideration as it can help shape different multi-embodiment configurations. Overall, while multi-embodiment configurations can complicate how identity is defined and perceived, they also offer valuable guidance for designing identity, especially within collaborative contexts.

2.4 The Need for Identity Design Principles

Anyone who creates a social agent is designing identity [187]. This includes creating personas using LLMs like ChatGPT, creating social robots, and all kinds of agents that are interacting with humans, including but not limited to embodied agents, chatbots, and conversational agents. Agents exhibit cues which provide signals to the user about the agent's identity; who it is, how it can be interacted with, how it should be treated socially, what kinds of social groups it belongs to, and what it might be able to do. The designer has control over *what cues* the agent has, depending on what identity design affordances are provided by the product or system. However, despite their best intentions, the designer has limited control over *how identity cues will be perceived and understood* by the user [90], or any human who might observe the agent and its interaction by simply being in the same environment [138, 140, 186]. Further, the identity designer may differ from the system designer [100, 218], complicating responsibility: while system designers cannot always control the identities others create, they remain accountable for harms that result. Thus, identity design must be consolidated with the end user's perspective in mind [98, 184].

A 2023 review identified only ten HRI conference papers defining themselves as work on "identity" [141], even though robot gendering and its implications have been studied and discussed for some time now within the community [65, 162, 163, 224]. It would appear that, whilst there have been some attempts to study the effects of

robot identity design, such as on agency [25] or trust [24, 25, 142], these have been rather focused on gender and lacking in intersectionality [141]. There have been fewer attempts to grapple with the what, how and why of robot identity design from a theory and methods perspective. This points to the difficulty of the task, and the need for design principles which can guide researchers, not least given ethical concerns regarding the "leveraging" of particular human social identity traits [194, 222] and the potential for influencing human-human norms and behaviours [194, 195, 222, 224].

Further difficulty is posed when artificial agents shift between embodiments, such as from a humanoid robot to a voice assistant, as they risk losing coherence in how they are perceived [12, 33, 79]. Without consistent identity cues, users may not recognise the agent as the same entity, leading to breakdowns in interaction, reduced trust, and disrupted collaboration [78]. These issues are compounded in multi-embodied systems, where agents must operate flexibly across time, modality, and context, often in complex social environments.

The risks associated with identity fragmentation are multifaceted. Functionally, discontinuity can trigger feelings of anger [203] or grief [73], break immersion [33], or impair team coordination and decision-making, especially when agents are embedded in high-stakes, distributed, or safety-critical settings. Cognitively, inconsistency in cues such as name, voice, behaviour, or role can not only confuse users, particularly children, older adults, or neurodivergent individuals who may rely on stability to interpret social presence, but can also engender feelings of eeriness known as the 'uncanny valley effect' [145, 148, 149]. Socially, poorly designed identities may reinforce harmful stereotypes, foster anthropomorphic misattributions, or undermine user agency by creating undue emotional or cognitive reliance on artificial systems [105, 230, 231]. These risks are increasingly being realised in AI systems, which are rapidly scaling without appropriate identity design guardrails in place.

To address these risks, identity must be considered as a multi-layered and intersectional construct operating across at least three levels: *individual, group*, and *societal*. We frame these as design lenses within our framework, rather than as types of agents, since different risks emerge at each level. These levels map key traditions in identity research: at the *individual* level the design goal is coherence of agent individual identity, drawing on concepts of personal identities and personas; the *group* level aligns with social identity theory's emphasis on group membership, role-based expectations, and entitativity considerations; and the *societal* level targets ethical accountability, representation, power, and inclusivity in technology design.

3 Developing the Framework

The ID framework was developed through a structured theory-driven synthesis of interdisciplinary literature in HCI, HRI, and related domains concerned with artificial agents, identity, and embodiment. Our aim was to address the under-constrained design problem of how to design coherent and interpretable identities for multi-embodied agents in collaborative contexts, and develop a transferable theoretical framework that may be applied to designing identity for artificial agents of any type.

The framework was developed iteratively. The first author initially developed a preliminary set of design principles for multi-embodied agents: a more complex case of identity design compared to single-agent systems. These initial principles were derived from a literature review examining the research on artificial agents, multi-embodiment, and artificial identity. These principles were then refined through a formal scoping review of the multi-embodiment research [32], which systematically examined how prior work has addressed embodiment transitions, identity cues, and identity continuity across HRI and related domains. Insights from the scoping review informed both revisions to the principles and the development of a set of design questions intended to support identity design for multi-embodiment systems.

The evolving framework, principles, and design questions were then refined in collaboration with the co-authors, comprising 14 experienced HRI and HCI researchers. These engagements served as a critical reflection process, enabling interrogation of the clarity, relevance, and scope of the emerging principles against broader research and design experience. Together, we clarified conceptual boundaries, examined the applicability of the principles across different agent types and interaction contexts, and reflected on the framework's applicability beyond a multi-embodiment setting. Through this process we strengthened the articulation and organisation of the framework. We present the resulting principles and guiding questions in this paper as the Identity Design Framework: a preliminary theoretical framework for supporting artificial identity design and evaluation activities.

This process also surfaced a set of open challenges and tensions in the design of artificial identity that extend beyond the scope of the framework itself. These include unresolved questions around identity continuity across embodiments, interactions, and temporalities, including how identity recognition might be operationalised and measured. Similarly there are outstanding questions around how identity can be effectively designed for multi-user settings and complex social dynamics whilst supporting plurality by design. Building on these insights, we co-developed a research agenda that outlines key challenges and directions for future work, including the application and evaluation of the framework, the development of methods for designing and assessing artificial identity, and the exploration of identity design across diverse agent types, contexts, and user groups. We present this research agenda as an invitation for the broader HCI and HRI community to engage with identity design as a shared design challenge.

4 The Identity Design Framework

To guide future development of artificial agent identities in socially embedded contexts, we introduce the *Identity Design Framework* for artificial agents, which structures identity design across three interrelated levels: **individual, group**, and **societal** (see Figure 2). The framework comprises twelve identity design principles and an accompanying set of design questions (see Table 1) intended to support the design, reflection, and evaluation of artificial agent identities across diverse interaction contexts.

Designers must consider multiple spheres of interaction when creating artificial agents, from interacting with individual users,

interactions with bystanders, groups, families, and teams, to interactions and implications for the wider communities, organisations, and societies into which such systems will be deployed and integrated. However, existing research offers limited support for helping designers systematically attend to the complexities and sensitivities of identity design across these contexts. Prior research in robot identity has distinguished between personal identity and social identity [102, 186, 187], while also drawing on social identity theory [93, 202] to examine social categorisation and group dynamics in human-agent interaction [187]. Most recently, Seaborn [187] defined three core identity types: *personal or individual* identities, *social* identities, and *superordinate* identities. At the same time, HRI and HCI explore the ethical risks and societal implications of artificial identity design, including concerns about misrepresentation, bias, power asymmetries, and individual and social harms [95, 141, 221, 222]. While these important theoretical works clarify how identity can be modelled, how it shapes social identity processes, and the socio-ethical implications of artificial identity, there is a gap in the research of practical methods that can guide identity designers during the design process. Our framework builds on this prior work, proposing that artificial identity should be considered across the following three interrelated levels or design lenses:

(1) At the **individual level**, identity distinguishes between specific artificial agents or personas based on attributes such as visual cues [31, 119, 138], behavioural consistency [11, 90, 109, 114], personality [11, 78, 173, 210], memory continuity [15, 203–205], and goal persistence [78, 107, 108, 189, 197].
(2) At the **group level**, identity design communicates the agent's role within a team [74, 75], social group [20–22], or community [102, 186, 194], supporting recognisable group membership, social alignment, and effective collaboration.
(3) At the **societal level**, identity design carries broader cultural meanings [61] and responsibilities, affecting how artificial agents reflect, reinforce, resist, or reshape social norms, stereotypes, systemic biases, and ethical expectations [140, 141, 160, 194, 222, 223].

4.1 Individual Level

Designing artificial agents at the individual level means considering what individual characteristics an artificial entity might have, including appearance, voice, name, social characteristics, behaviours and other identity cues it exhibits to signal to users its individual identity. While this is similar to the construction of *personal identity* [102, 187] in humans, we avoid using the term *personal* for artificial systems. It is a deliberately designed construct, created either by a human or a generative system, that presents the appearance of individuality without implying subjectivity or moral status. For this reason, we use the term *individual identity* to describe the level at which one artificial agent can be distinguished from another.

Early work on individual identity often used multi-embodiment as a high-visibility testbed, examining how one agent maintains a coherent sense of self across media and embodiments [32, 138, 158]. In such high-variation settings, the requirements for a coherent individual identity become especially salient. For example, Martin et al. [138] explored which visual cues help people recognise a migrating virtual agent. The exhibition of continuous traits across forms, such as personality [12, 78], memory [15, 203], intent [78, 107], and communicative style [22], can help to establish a stable and intelligible agent identity in people's minds. Without this, users may perceive each embodiment as a different agent, reducing trust and degrading the relationship. Laity et al. [119] propose a framework for decomposing and reconstructing a robot's persona across embodiments, based on visual, auditory, and behavioural identity and migration signals. While they have systematically reviewed the identity cues in the multi-embodiment research, they have not tied in how these cues are linked to a theoretical model of artificial identities, nor is there an existing set of guiding principles for the design of such agents. We thus propose to design social agents at the individual level along five principles: **recognisability**, **consistent behaviour**, **identity continuity**, **memory**, and **persistent goals**.

4.1.1 Recognisability. Recognisability is foundational at the individual level: users should be able to identify the same agent across channels, platforms, and embodiments [15, 138]. Miranda et al. [141] defined fourteen axes of human-like identity (such as age, class, disability, ethnicity, gender, nationality, and more) and found that there has been limited operationalisation and intersectionality of identity overall in HRI research. In multi-embodiment research, designers have proposed a range of *identity cues* to support identity recognition, to varying levels of success [31, 119]. Identity cues used include name [20, 159], voice [107, 203], colour palette [138], sound design [78], and behavioural style [10, 173].

While some studies show that users can recognise migrating agents based on these cues [138, 173], findings are inconsistent. Recognition often improves with repeated interactions [108], but short-term studies show mixed results, particularly when multiple cues are used together, making it difficult to isolate the effects of each. For example, Cuba [47] found personality to be the most influential cue, followed by visual appearance, while Gomes et al. [79] reported that many users failed to recognise a migrating pet agent, despite consistent behaviours. Few studies clearly define how recognition is measured. Many rely on subjective judgements or hypothetical tasks rather than behavioural indicators. This makes it difficult to generalise findings to real-world applications, where users may not be explicitly told that migration is occurring, or how many agents are present.

4.1.2 Consistent Behaviour. To support individual identity across contexts, agents should exhibit persistent patterns of behaviour with coherence under bounded adaptation [12, 78, 109, 147]. People might draw on cues such as tone of voice [22], gestures, speech rhythms, and affective responses [109] to interpret an agent as a unified social entity. Even subtle regularities, like speech pauses or gestures, could signal continuity across contexts and embodiments.

A stable personality further reinforces identity [11, 78, 138, 172, 173, 211]. Traits such as formality, empathy, or humour should remain consistent across forms to foster user trust, familiarity, and a sense of coherence over time. Prior work has shown that users appreciate agents with stable personalities and often desire to customise these traits to align with their own preferences [172].

Behavioural consistency has been shown to aid or hinder recognition of migrating agents [12, 78]. Inconsistencies, whether due to technical limitations or design, can disrupt the illusion of continuity. For instance, users in Gomes et al. [78] perceived a virtual and

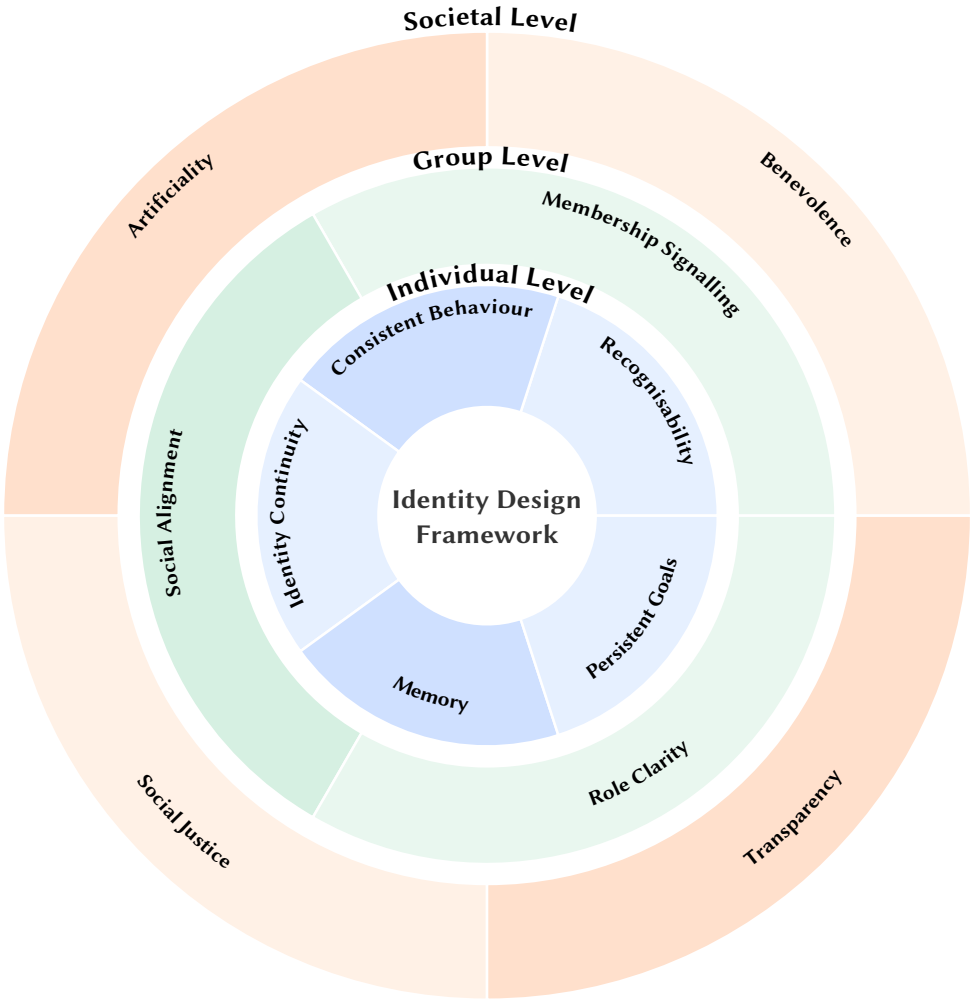

Figure 2: The Identity Design Framework for Artificial Agents

physical dinosaur agent as separate entities due to a time lag during the migration process, even though the agent's goals and memory were preserved. Similarly, Segura et al. [189] found that although agent needs and goals migrated across forms, differences in how the agent pursued those goals undermined identity coherence.

In comparative studies, extroverted or distinctive behavioural styles were more reliably recognised than subtler, introverted ones [11]. This suggests that salient and consistent behaviours, paired with other cues, play a crucial role in helping users maintain a stable mental model of the agent across embodiment shifts.

4.1.3 Identity Continuity. Maintaining continuity of identity and perceived entitativity is particularly important in multi-embodied agents due to their capacity to assume multiple forms. Designers must consider the intended model of social presence for the system, whether it involves re-embodiment, co-embodiment, or all-for-one identity configurations [135]. To support a coherent sense of identity, migration cues play a critical role in helping users mentally link different embodiments as representations of the same agent [15, 32, 78, 119]. Linguistic references and pronoun use can further reinforce entitativity. The way multi-embodied agents or agent collectives speak about themselves provides additional cues to continuity. For example, consistent use of first-person pronouns across embodiments or coordinated group reference styles can help users perceive a unified identity across forms [21, 22]. Poor migration signals can lead to confusion about the embodiments of an identity and may obscure how users assign identity to each form [119]. Thus, effective migration cues should clearly communicate both the migration process and its outcome, even for first-time users. Zhu et al. [232] propose four stages of migration that can help structure this process and inform the design of transitions across embodiments: connect, connecting, robot connected, and robot appearing.

4.1.4 Memory. Continuity of memory matters across channels, sessions, and embodiments. Users should not need to reintroduce themselves or repeatedly provide the same information each time an agent reappears in a new form. As Aylett et al. [15] and Tejwani et al. [203, 205] emphasise, shared memory is essential for enabling

Table 1: Identity Design Principles and Questions for Artificial Identity

Level	Design Principle	Design Questions
Individual	1) Recognisability	What identity cues enable users to recognise an agent as distinct from other agents in the system? What social presence model is the identity based on, and does this make sense to users?
	2) Consistent Behaviour	What personality traits, behaviours, and affect does the agent consistently perform across embodiments, environments, and time? When does the agent behave differently in different forms, and what effect does this have on the perceived identity?
	3) Identity Continuity	What migration cues enable users to understand and perceive the continuity of the entity across embodiments and interactions?
	4) Memory	What aspects of an agent's memory (information and history) are available across embodiments, environments and sessions? How might this information shape the agent's personality and behaviour? What are the consequences of storing or deleting this information?
	5) Persistent Goals	What goals, tasks, and intents persist across agents' embodiments to support users?
Group	6) Membership Signalling	What cues enable users to recognise an agent as a group, team, or community member?
	7) Social Alignment	What goals and behaviour does the agent show to align with group, team, or community goals and social norms? How might the agent's social alignment privilege or exclude different social groups?
	8) Role Clarity	What is the agent's role within the group, team, or community context? Is this clear and interpretable to the user? How might the agent's roles differ across different users and user groups?
Societal	9) Benevolence	What features protect users from being deceived and manipulated? How might we ensure the agent's identity is benevolent and prevents individual or social harm?
	10) Artificiality	What cues signal the artificial nature of the agent to distinguish it from humans? How might the identity design hide or reveal the artificial nature of the agent?
	11) Social Justice	What features of the design might reinforce harmful stereotypes or systemic bias based on marginalised identities? How might the identity design contribute towards social justice?
	12) Transparency	What aspects of the identity design obscure what the system does, how reliable it is, or how it works? How might we make the agent's mode of control (e.g. WoZ or LLM), capabilities, limitations, and embodiment changes clear to users?

seamless agent migration. Agents that retain knowledge of prior interactions, user preferences, and contextual history are better able to deliver personalised and coherent experiences, reinforcing the perception of a continuous, unified identity across embodiments. However, prior work also cautions against assuming that continuity of memory is always desirable. Reig et al. [171] found that users' expectations for memory retention vary depending on the context. For instance, people may welcome personalised service in private or professional settings, e.g., healthcare, but find it intrusive in public or commercial contexts like department stores. Concerns also arise around privacy and social appropriateness, particularly when sensitive information is spoken aloud in shared spaces where others may overhear [171]. Furthermore, users may object to cross-embodiment memory sharing when the new form is not perceived as the same agent; Tejwani et al. [203] found that participants viewed the sharing of information across distinct embodiments without consent as a violation of privacy and trust.

4.1.5 Persistent Goals. Continued intent communication is necessary to strengthen the perception of the agent as a coherent, goal-directed social presence [91] and is particularly important for fostering trust in collaborative or assistive settings. Agents should show continuity of intent, maintaining stable goals and tasks across channels, platforms, and embodiments, while adapting their enactment to the affordances of each context. If an agent appears to pursue different aims or behaves inconsistently across embodiments, users may perceive it as a different entity [10, 78], which could undermine collaboration and increase cognitive load.

However, agents embodied in different forms may require different goals or interaction modalities. Designers should therefore carefully consider which goals, tasks, and intentions should persist across embodiments, and which should adapt to the constraints and affordances of each form and environment. For example, Segura et al. [189] examined the migration of agent needs and internal states, but the agent's behaviour was adapted across physical and virtual embodiments to suit the capabilities of each form. Task continuity has also been investigated in prior work [107, 108, 189, 197], highlighting its importance for supporting seamless interaction across embodiment transitions.

4.2 Group Level

While individual identity design is critical, agents also function within collectives, like groups, teams, organisations, and communities, where identities must be designed and interpreted in relation to others [22, 75, 186, 194]. Group identity might involve signalling group membership or entitativity, aligning with shared goals and norms in collaborations, or performing recognisable roles. A clear group identity should help establish shared mental models, clarify expectations, and support coordinated action across agents and humans. It can also enhance perceptions of system coherence, especially when multiple robots or agents are present [21, 22].

The Social Identity Approach (SIA) [201, 208] provides a useful framework for understanding how people categorise others, and might categorise artificial agents, based on perceived group membership [186, 194]. As such, agents might be designed to align not only with functional roles but also with broader social group identities, e.g., gender, cultural background, or professional role, where appropriate. Providing the ability to customise and personalise identities to suit user preferences may also help to achieve closer alignment with users' social groups [172, 194]. This alignment may be achieved through cues in appearance, language, behaviour, or values, but must be handled sensitively to avoid stereotyping or reinforcing bias [95, 140, 141, 194, 221].

Designing for group identity also involves considering how groups of users will interact with agents, whether as active or passive users [99, 227]. Voice assistants like Google, Alexa, or Siri are often used in family or group settings [122, 165], while robots deployed in homes, shopping centres [57], airports [207], or other public spaces must also interact with multiple people simultaneously. These types of settings raise challenges of coordination, synchronisation, and group dynamics [177], requiring systems to manage competing inputs while maintaining coherent identities at the group level.

Furthermore, multi-embodiment and multi-robot contexts add further complexity as the design of multiple agents and associated embodiments can critically impact user perceptions and interactions. For instance, multi-embodiment configurations implemented through different group identity strategies can shape perceptions of trust [217] and entitativity [21, 22]. These perceptions in turn affect user expectations and interactions, and can vary across cultures [70]. Additionally, the number and type of robots present in groups can influence user attitudes and emotional responses [71]. As such, in these contexts, coherence and role clarity become especially important to prevent confusion and help manage user expectations and perceptions. We propose three principles that form the foundation of designing identity at the group level: **membership signalling**, **social alignment**, and **role clarity**.

4.2.1 Membership Signalling. Clear group membership is essential for effective agents in group contexts. Agents can be designed to align with specific human social groups (like gender, profession, or age), or with other agents (signalling common functionality, swarm membership, or affiliation with a particular company) [187]. For example, in human-robot teaming contexts, clear signals of group membership reinforce affiliation, support coordination, and strengthen perceptions of the agent as a legitimate team member. Identity should be consistently tied to the team through branding, visual design, language use, and adherence to behavioural norms. Robots are more likely to be perceived as teammates when visually marked as group members, like wearing uniforms or adopting shared colours [70, 72, 75]. In multi-agent systems, especially those with heterogeneous embodiments, affiliation must be clearly communicated across form and context. Designing for group membership also requires attention to *entitativity* in multi-robot groups or environments with multiple agents. Robot groups may be mentally modelled differently than human groups [20]. Shared behaviour and qualities can increase perceived entitativity, while unique observables encourage one-for-one presence models. In multi-robot groups, users may rely on cues such as who is speaking, how agents self-refer and refer to other agents, names, and distinctiveness of names and voice [20]. But, multiple cues may be needed for users to perceive differences between group members [22].

4.2.2 Social Alignment. Where agents are designed as members of social groups, they should act consistently with the group norms [89, 172], expectations, and goals [27, 46, 121, 180]. This includes adopting contextually appropriate interaction styles and cooperating with humans and other agents. Behaviour that deviates from norms through excessive autonomy or unpredictability risks the agent being perceived as an outsider [187], eroding trust and reducing willingness to collaborate. Persistent alignment with group values and practices helps reinforce the agent as a stable group member rather than a transient participant or unstable tool. Where people outnumber robots, stronger group affiliation cues might be required, as large groups of humans can be less inclusive of robots [219]. To be accepted as genuine group members, agents must also contribute meaningfully to collective goals. Their actions should not be perceived as isolated or self-serving, but as supporting broader team objectives [40, 188]. In high-stakes contexts such as search-and-rescue, emergency response, or medical collaboration, users need to clearly interpret how each agent's actions support mission success; agents should maintain stable goals and communicative strategies across forms, aligning with overarching aims. Fragmented or inconsistent contributions could confuse users, create redundancy, and undermine collaboration.

4.2.3 Role Clarity. Role clarity is critical for agents operating in group and collaborative contexts, as roles set expectations about capability and behaviour. Clearly defined roles help users understand what tasks an agent is designed to perform, the limits of its competence and authority, and how they should engage with it [29]. Expectations differ substantially across roles and contexts; for example, users may interpret a humanoid home robot differently from a robot receptionist, or a service desk agent differently from a general-purpose virtual assistant. In multi-user settings, an agent's role may also vary across users. In domestic environments, for instance, a robot may adopt a supportive or service role for adult users, supervisory or educational roles for children, and caretaker roles when interacting with pets. Thus, in family-robot interactions, it is important to clearly define a robot's roles when interacting with different members to enhance both individual interactions and family connections [39, 199].

Role clarity becomes particularly important for multi-embodied agents, which may transition across forms, environments, or interaction modalities while maintaining a shared identity. If an agent

acts as a scout in one form and an advisor in another, users may struggle to track an agent's function within the team, undermining their ability to form stable mental models of the agent. Prior work suggests that agents re-embodying across domains that imply different levels or types of expertise can be perceived negatively [171], indicating that role transitions must be designed with care. Similarly, when multiple agents share embodiments, this can also blur boundaries between different agent identities [159, 170, 217]. As such, consistency in task-related behaviour and role-based cues, such as distinctive speech registers, interaction modalities, explicit communication of roles through uniforms, names, and verbal instructions, or form-specific capabilities, helps users form mental models of the agent's purpose and expertise. In turn, this facilitates smoother coordination, clearer expectations, and more effective division of labour in collaborative and group contexts.

4.3 Societal Level

The identities of social agents are never neutral. They are embedded in broader social [187], political [226], and cultural contexts [61], therefore the design of social agents must be approached with care due to the ethical and social risks they present [141, 187, 215, 222]. The societal level of identity refers to how the agent's design reflects, reproduces, or resists social norms, values, and power structures more broadly within the wider social sphere. Without critical attention, agent identities can reproduce stereotypes, marginalise certain groups, or enable deceptive design practices (e.g., creating artificially "friendly" personas that exploit trust or emotional vulnerability) [64]. Identity design must therefore be grounded in commitments to transparency, accountability, and social justice.

Recent advances in generative AI and large language models make it technically feasible to craft agents that closely mimic human traits like voices, personalities, visual styles, and even backstories [14]. These capabilities expand opportunities for personalisation but also raise significant risks, such as female-gendered chatbots or robots in service roles reinforcing occupational and cultural stereotypes [5]. AI-enabled robots amplify these risks as they interact with real-world environments, where identity cues may not only persuade but also enable physical actions that increase trust, dependency, or harm. Deepfake technologies and AI-driven scams further illustrate how identity cues can be weaponised for deception [150] by creating realistic artificial identities to extract information, money, or influence. A recent Danish proposal [157] to extend copyright protections to identity characteristics illustrates growing societal unease with the unauthorised reproduction of human identity. Such risks highlight the need for design practices that protect individuals and communities from identity theft, reputational harm, technology-enabled abuse, and psychological manipulation [225].

Identity at the societal level is temporal; through a balance of continuity and change that links past, present, and future [23], artificial identities thus cannot be treated as static. Design choices evolve through system updates, retraining, or user adaptations, and their meaning shifts as cultural norms change. An identity feature considered inclusive today may be seen as stereotypical tomorrow. Artificial identities can enable people to converse with deceased loved ones and process grief [62, 110], provide a medium for digital immortality [181], or assist them to imagine their future selves [101]. Whether the benefits of this outweigh the risks and associated harms remains unclear. These issues are already being explored in popular culture, like Black Mirror episode *Be Right Back* [37], where an AI copies the identity of a dead partner. While comforting at first, it becomes disturbing over time as the line between real memory and artificial imitation breaks down. The challenge for designers lies in anticipating identity evolution, supporting how artificial identities will evolve over time, and in creating mechanisms that maintain continuity and accountability.

Identity is shaped by culture and context [178, 185, 200, 228]. In multicultural contexts, individuals navigate a repertoire of identities [92]; as societies become more interconnected, social identities grow more diverse [174]. Who we are, and how we express that identity, shifts depending on where we are, who we are with, and what roles are salient in that moment [192]. We propose four principles for designing artificial identities at the societal level: **benevolence**, **artificiality**, **social justice**, and **transparency**.

4.3.1 Benevolence. Benevolence is an imperative that agents should not be designed to intentionally deceive or manipulate users into behaviours that are not in their best interests. This includes avoiding embedding "dark patterns" [34, 35] through subtle deceptive identity cues that exploit user trust or vulnerability [64]. Deceptive cues may include overly anthropomorphic traits to trick users into believing agents have more humanlike capabilities than they actually do, emotionally charged backstories which could lead users to actions they would not have intended to do such as buying a product they do not want or need, or falsely implied competencies that might encourage users to overtrust agents or purchase a robot that has less features than it appears to. For example, designing an agent to appear as a trusted authority figure (like a teacher, doctor, or friend) without the appropriate safeguards or transparency may lead users to over-disclose information or comply with suggestions they would otherwise reject [113]. Ethically informed identity design must therefore centre on user welfare, avoiding manipulation, coercion, or emotional exploitation.

4.3.2 Artificiality. Artificiality refers to signalling an agent's non-human nature. Human-like form and fluent dialogue can mask an agent's limitations, inflating perceptions of intelligence and autonomy [81, 111]. This can lead people to misinterpret agents as human, attribute them undue moral capability, misplace trust, or form problematic attachments [55, 85, 88]. Fictional or media portrayals of robots can widen the expectation gap, increase over-attributions, and undermine trust and acceptance [94, 175, 176]. Embodiment and expressive behaviours also amplify perceptions of emotion and intentionality [26, 149]. Children, older adults, or people with cognitive impairments may be especially vulnerable to these effects [36, 54, 63]. Artificiality cues like synthetic voices, restrained anthropomorphism, and explicit disclosures of system limitations calibrate expectations [52, 222]. Designers can embrace the role of "Uncanny Killjoy" and advocate for artificiality to reduce anthropomorphic design risks [187]. Under the "honest signals" view [90, 97], no sociability cue exhibited by an agent is entirely truthful, so designers must balance usefulness against the risks of deception. A practical example is disclosure-based prompting, such as an agent responding to "How are you feeling?" with "Although I have no emotions, my systems are functioning as intended", which

preserves sociability while clarifying identity [111]. Artificiality cues mitigate risks of reinforcing stereotypes or concealing political design choices, supporting accountable deployment [52, 187].

4.3.3 Social Justice. Social justice addresses the broader social and cultural implications of identity design. The names, voices, appearances, accents, and behavioural traits embedded in artificial agents are never neutral [140, 222, 226]. These cues reflect (and can reinforce) cultural norms, social hierarchies, and systemic biases. Without careful attention, design choices may unintentionally reproduce harmful stereotypes or marginalise particular groups. For instance, consistently assigning certain ethnic or gendered voices to service roles, or exoticising specific cultural markers in persona design, can perpetuate inequities rather than challenge them [5, 141, 216]. Yet, diversity in identity design and the leveraging of stereotype-breaking behaviours might be beneficial in the context of diverse classrooms [142] or challenging inappropriate user behaviour [224]. Social justice requires designers to reflect critically on whose identities are represented, how, and to what ends. Designers must ask "what does the identity design afford, and for whom?" [49, 50]. This includes questioning default design choices, engaging with diverse users and stakeholders, and foregrounding equity, inclusivity, and accountability in all stages of identity development.

4.3.4 Transparency. Systems should be upfront about how they work and who is accountable for their functioning [68, 84]. Transparency helps humans and agents align in goals, processes, and intent [9, 132, 136, 212]. For identity design, this means communicating how autonomous an agent is, who controls it, who designed it, and for what purpose. Transparency can be achieved through explicability or verbal explanations, as well as through multimodal cues such as tone or facial expressions, which influence how intent is understood [168, 206]. Two challenges are central: *calibration* and *information overload*. Transparency can reduce over- or under-reliance, while its absence risks deception [64, 104, 151, 182]. Yet excessive cues, such as too many screens or competing modalities, can overwhelm users, with social signals like robot gaze drawing attention away from information presented on accompanying screens [126]. Practical strategies include identity disclosures that clarify autonomy and design intent. Systems might answer "who controls you?" with information about human oversight, or take a normative step and provide moral transparency by disagreeing with harmful beliefs (like climate change denial) [52, 214]. Transparent identity cues should also consider cultural and institutional contexts [169, 187], making transparency a socio-emotional, as well as a technical, design challenge.

5 Discussion and Research Agenda

In this work, we introduced the Identity Design (ID) Framework, a layered framework for building coherent, interpretable identities in artificial agents as they populate both experimental and commercial human-robot interaction and artificial intelligence systems. The framework consists of 12 identity design principles, organised across three levels of interaction, and a set of design questions, and is intended to guide designers to develop artificial identities and identity systems responsibly. By viewing identity through and across individual, group, and societal lenses, we argue that designers can safeguard usability and trust while reducing risks of manipulation, confusion, and harm. At the individual level, the principles focus on recognisability, behavioural consistency, continuity, memory, and persistent goals; at the group level, they situate agents within teams and shared roles; and at the societal level, they foreground cultural and ethical responsibilities. In the sections that follow, we provide guidance on using the framework, introduce a research agenda for advancing the field of identity design, and discuss limitations and future work.

5.1 Guidance on Using the Framework

Rather than prescribing fixed rules, the Identity Design Principles are formulated as guiding questions to support reflection during the ideation, prototyping, and evaluation of artificial identities, as well as in the development of systems that enable user-generated identities. The design questions are intended to prompt designers to consider how identity is constructed, communicated, and interpreted across different interactions, embodiments, and contexts. The framework is not a checklist to be applied uniformly, but a flexible tool that can be adapted to specific design goals, technological capabilities, and user populations.

We intend these principles to function both as design prompts and evaluative criteria: teams can reflect on the guiding questions during ideation, use them to structure embodiment transitions, and apply them as a deployment checklist. They are not prescriptive but an evolving framework to be adapted to specific contexts, technologies, and user groups, and integrated with adjacent concerns in affective computing, accessibility, and responsible AI. Participatory methods and interdisciplinary collaboration remain essential to ensure identity design is not only technically sound but also socially and culturally responsive.

The three levels of the framework can be applied to any type of artificial agent. They operate as design lenses that direct attention toward progressively wider spheres of interaction, beginning with the individual agent, expanding to teams and multi-agent systems, and extending to the organisations, communities, and social groups in which users are situated. This layered structure is intended to support designers in attending to individual identity cues, group dynamics, and broader societal implications in an integrated way.

At the same time, these levels are analytically distinct but practically entangled. In practice, identity designs often affect multiple levels. A single cue may support recognisability at the individual level, signal affiliation at the group level, and reinforce or challenge stereotypes at the societal level. We therefore encourage designers to use the framework relationally, tracing how identity cues travel across levels and interaction contexts. In this regard, intersectionality functions not as a concern confined to the societal level, but as a cross-cutting analytical lens [141, 183, 223]. It enables designers to examine how multiple identity axes interact across all three levels, and how their effects may compound across users and contexts.

The principles are intended to complement existing approaches to system design. This includes fundamental interaction design principles [156], universal design guidelines [45], categorisation and analysis of communication [90], and principles of responsible robotics [30]. Used alongside these broader frameworks, the Identity

Design Framework can help designers and researchers evaluate whether artificial identities remain coherent, interpretable, and ethically grounded as agents move between forms and contexts.

5.2 Research Agenda

The principles and framework presented in this work provide an important conceptual foundation for designing artificial agent identity, but they remain preliminary and require further work to examine their applicability to artificial agent design more broadly. Research on identity design for artificial agents is still in its early stages, and the field would benefit from sustained empirical, methodological, and ethical development.

To address these gaps, we propose the following research agenda that outlines key directions for advancing the design of artificial identity. This includes developing approaches for the operationalisation and measurement of identity, extending research beyond single-agent paradigms, embedding risk assessment and plurality by design, examining identity design as a distributed practice, integrating speculative and performative design methods into design practice, and examining the role of automation in shaping and maintaining artificial identity. We argue that artificial identity is an ongoing design challenge, requiring new methods, evaluation approaches, and critical perspectives to support its development in practice. We present these directions in the following sections.

5.2.1 Strengthen operationalisation and measurement. The operationalisation and measurement of identity in artificial agents remain underexplored, with little understanding of how identity is recognised and perceived by users. Studies should develop robust behavioural recognition tasks (e.g., forced-choice matching under time pressure), manipulate single versus combined cues in pre-registered factorial designs (e.g., [143]), and run longitudinal sessions to capture learning effects (e.g., [3, 115, 117, 118]); given mixed effectiveness of recognisability cues [82, 128], researchers should systematically vary cue salience and test it with personality stability and communicative style (e.g., [164]), and standardise migration manipulations with structured transition stages and explicit linguistic signals of entitativity (e.g., first-person continuity) while evaluating additive effects on recognition, trust, and task continuity.

5.2.2 Move beyond single-user paradigms. Identity research has largely focused on single human-agent interactions, leaving group identity and identity at scale underexplored. Yet many contexts involve agents working with multiple users, including home, team, organisation, and multi-agent environments. More research is needed to examine robot identity in these contexts, manipulating/measuring group membership (see [155]), alignment with norms and goals, and role clarity across heterogeneous embodiments (e.g., branded vs. unbranded agents; stable vs. shifting roles). Addressing these dynamics will generate insights into how artificial identities are negotiated in real-world settings and inform design strategies that support effective teamwork and ethical deployment.

5.2.3 Embed risk assessment and plurality by design. Research on identity at the societal level has so far been limited, often focusing on issues like gender representation or trust while leaving intersecting axes of identity largely unexamined [141, 183, 223]. Yet identity choices can risk reinforcing stereotypes, obscuring accountability, or introducing new inequities. To mitigate these risks, toolkits should embed fairness audits (e.g., [1]), paradigm evaluation when administered via robots (e.g., [2]), "identity handoff" tests for embodiment changes (e.g., [114]), and frameworks that probe deception or over-personalisation (see [64]); packaging these as preregistered open protocols, code, and datasets can turn identity design from a craft into an accumulative, testable science. Guidance also comes from feminist HRI's reflexive questions [223], matrix-guided technology power analysis for interrogating the "power landscape" of robotics [216], and ethical risk assessment templates for anthropomorphic robots [221, 225]. Together, these directions support identity design that is equitable, intersectional, transparent, and accountable.

5.2.4 Examine identity design as distributed practice. The design of artificial identity is no longer confined to researchers, designers, or engineers. Users are now actively involved in creating social agents through customising prompts, personalising behaviours and curating and sharing agent personas over time in platforms like Character.AI or Replika [28, 80, 231]. Identity design is thus emerging as a distributed practice spanning system developers, designers, and end-users; each of whom may uphold or violate the identity principles in different ways. Designers may engage with identity design reflectively and strategically, seeking to influence social change, or may weaponise artificial identity for political or commercial ends. Everyday users may prioritise convenience or familiarity [80], overlooking the implications of their choices. They may also engage playfully to explore their own personal identities [28], or create and rework artificial identities to address social injustice and algorithmic bias embedded by structural inequalities and the companies that produce these systems [38, 141, 223]. This raises questions about how identity design principles translate into practice across different stakeholder groups; how designers interpret and operationalise them, how platform affordances shape or constrain their application, and how everyday users engage with, overlook, or actively contest the identity choices embedded in the systems they use.

5.2.5 Integrate speculative and performative design methods into design practice. Identity can be performative, philosophical, and political, as well as an outlet for creative expression [7, 17, 86, 127]. Speculative and performative approaches offer powerful ways to explore artificial identity beyond functional requirements and lab studies, yet remain underused in the development of robotics and conversational AI systems. We can imagine new artificial identities that deliberately "escape" current norms and practices through speculative design, e.g., exploring alternative futures [59, 60, 220].

Increasingly, robots appear on stage [6], where theatre functions as an artistic platform and as a testbed for HRI: not to mimic human behaviour, but to use theatrical context, interpretation, and progression to probe new modes of interaction. In this spirit, the stage is more than a rehearsal room [198]; it is a space to examine identity as a performative act. Theatre and dance methods have been explored, such as movement [16] (also with drones [56]), acting systems [87, 106], improvisation [8, 120], and participatory design in assistive technologies [154]. The illusion of an identity depends

on the audience's willingness to suspend disbelief [58]. Unlike puppetry, where the puppeteer's presence makes the artifice explicit, robots are presented as autonomously animated, and designers and operators recede from view while audiences are invited to accept the performance as "alive". Such approaches, or new ones inspired by them, should be integrated into artificial agent design processes. This will enable researchers and practitioners to probe identity as both an engineered construct and a lived phenomenon, opening a poetic space for creative and accountable identity design.

5.2.6 Towards automating identity. Recent advances in generative AI (like LLMs and diffusion models) enable us to endow agents and robots with distinct identities. At the individual level, progress has been made in persona generation, long-term memory, goal-directed behaviour, coherent conversational patterns, and audio-visual cue generation [42, 103, 133, 190]. Yet a gap remains in embodying identity through consistent non-verbal behaviours such as co-speech gestures, backchanneling, and motion patterns, where current methods achieve only partial success [229].

Another unexplored frontier is automated generation of behavioural cues that persist across diverse physical or virtual forms (heterogeneous embodiments; e.g., [51]). At the group level, while role assignment and norm-following can be implemented through in-context learning and fine-tuning of LLMs [96, 130], future research can look at how agents can actively express group membership, adapt to dynamic roles, and interact fluidly with humans (e.g., [77, 129, 161]).

At the societal level, technical solutions alone are insufficient; consensus-driven guidelines and transparent integration into generative systems are essential to ensure that identity cues are fair, trustworthy, and culturally appropriate [166]. Systematic benchmarking is also needed to evaluate the validity and consistency of identity realisation, extending beyond linguistic persona fidelity to multimodal and group-level effects (e.g., [179]). Ultimately, research must integrate technical feasibility with benchmarking, social responsiveness, and normative alignment if the aim is to transform identity design from ad hoc practice into a cumulative, scientifically grounded discipline.

5.3 Limitations and Future Work

This work is not without limitations. The ID framework was developed through theoretical synthesis of the research within a multi-embodiment context, and its applicability beyond this domain, such as to single-embodied, non-collaborative, or long-term social agents, remains to be established. In abstracting principles from situated design cases, the framework necessarily foregrounds recurring identity tensions while omitting some contextual nuance; its application therefore relies on designer interpretation rather than providing prescriptive guidance. The framework also does not yet specify how the principles can be operationalised in practice, for instance, how they might be translated into concrete design heuristics, evaluation rubrics, or tools that support their application across different design workflows and agent configurations.

The principles presented here are preliminary and have not yet undergone a formal evaluation in applied design settings. Future work will involve structured co-design and application-led studies to assess the clarity, completeness, and practical utility of the framework across diverse domains and agent configurations. Further work could also examine how the principles can be applied to existing artificial agent systems, or when identity tensions arise dynamically during deployment.

6 Conclusions

Artificial identity is a core design concern in HRI and HCI that shapes usability, trust, and the broader social impact of artificial agents. The Identity Design Framework, comprising twelve principles that span individual, group, and societal levels, can guide designers, engineers, and researchers in creating coherent and interpretable identities across robotic, virtual, and multi-embodied agents. We outline a research agenda to 1) operationalise and measure identity, 2) move beyond single-user paradigms to groups and multi-agent settings, 3) embed risk assessment and plurality by design, 4) examine identity design as distributed practice, 5) fold speculative methods into design practice, and 6) advance techniques for identity automation. Treating identity as an explicit design parameter can yield agents that are more coherent and effective. Our contribution is a shared language and a practical scaffold for designing artificial identities that are recognisable and continuous for individuals, legible and coordinated in groups, and transparent, equitable, and benevolent at the societal level. We invite the community to refine, extend, and audit these principles across long-term deployments and diverse cultures.

Acknowledgments

This project was supported by the Commonwealth of Australia, as represented by the Defence Science and Technology Group of the Department of Defence. This work was supported by the Engineering and Physical Sciences Research Council [grant numbers EP/V00784X/1, EP/Y009800/1].

References

[1] Nida Itrat Abbasi, Fethiye Irmak Dogan, Guy Laban, Joanna Anderson, Tamsin Ford, Peter B. Jones, and Hatice Gunes. 2025. Robot-Led Vision Language Model Wellbeing Assessment of Children. In *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Eindhoven, Netherlands, 59–64. doi:10.1109/RO-MAN63969.2025.11217833

[2] Nida Itrat Abbasi, Guy Laban, Tamsin Ford, Peter B. Jones, and Hatice Gunes. 2024. Robotising Psychometrics: Validating Wellbeing Assessment Tools in Child-Robot Interactions. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, Pasadena, CA, USA, 1651–1658. doi:10.1109/RO-MAN60168.2024.10731253

[3] Nida Itrat Abbasi, Guy Laban, Tamsin Ford, Peter B. Jones, and Hatice Gunes. 2025. A Longitudinal Study of Child Wellbeing Assessment via Online Interactions with a Social Robot. *ACM Transactions on Human-Robot Interaction* 14, 3 (June 2025), 1–35. doi:10.1145/3722123 GSCC: 0000005 2026-01-14T04:45:15.401Z 0.17.

[4] Anna M. H. Abrams and Astrid M. Rosenthal-von der Pütten. 2020. I–C–E Framework: Concepts for Group Dynamics Research in Human-Robot Interaction. *International Journal of Social Robotics* 12, 6 (Dec. 2020), 1213–1229. doi:10.1007/s12369-020-00642-z

[5] Neziha Akalin, Maria Arnelid, and Katherine Harrison. 2025. Gendering Robots in Human-Robot Interaction: An Interdisciplinary Approach. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE Press, Melbourne, Australia, 1104–1110. doi:10.1109/HRI61500.2025.10974049

[6] Irene Alcubilla Troughton. 2025. *Moving Together: A Performing Arts Approach to Relational Human-Robot Interaction Design.* Doctoral Thesis. Utrecht University, Utrecht. doi:10.33540/2700

[7] Irene Alcubilla Troughton, Kim Baraka, Koen Hindriks, and Maaike Bleeker. 2022. Robotic Improvisers: Rule-Based Improvisation and Emergent Behaviour in HRI. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Sapporo, Japan, 561–569. doi:10.1109/HRI53351.2022.9889624

[8] Irene Alcubilla Troughton, Hendrik von Kentzinsky, Maaike Bleeker, and Kim Baraka. 2023. "Improvisation ≠ Randomness": a Study on Playful Rule-Based Human-Robot Interactions. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Busan, Korea, Republic of, 52–59. doi:10.1109/RO-MAN57019.2023.10309523

[9] Georgios Angelopoulos, Mehdi Hellou, Samuele Vinanzi, Alessandra Rossi, Silvia Rossi, and Angelo Cangelosi. 2025. Robot, Did You Read My Mind? Modelling Human Mental States to Facilitate Transparency and Mitigate False Beliefs in Human–Robot Collaboration. *ACM Transactions on Human-Robot Interaction* 15, 1 (Aug. 2025), 1:1–1:29. doi:10.1145/3737890

[10] Krzysztof Arent and Bogdan Kreczmer. 2013. Identity of a Companion, Migrating between Robots without Common Communication Modalities: Initial Results of VHRI Study. In *2013 18th International Conference on Methods & Models in Automation & Robotics (MMAR)*. IEEE, Międzyzdroje, Poland, 109–114. doi:10.1109/MMAR.2013.6669890

[11] Krzysztof Arent, Bogdan Kreczmer, and Łukasz Małek. 2011. Identity of Socially Interactive Robotic Twins: Initial Results of VHRI Study. In *2011 16th International Conference on Methods & Models in Automation & Robotics*. IEEE, Międzyzdroje, Poland, 381–386. doi:10.1109/MMAR.2011.6031377

[12] Krzysztof Arent, Bogdan Kreczmer, and Łukasz Małek. 2012. Identity of a Companion, Migrating between Robots Significantly Different in Terms of Expressive Capabilities: Initial Results of VHRI Study. In *2012 17th International Conference on Methods & Models in Automation & Robotics (MMAR)*. IEEE, Międzyzdroje, Poland, 262–267. doi:10.1109/MMAR.2012.6347877

[13] Ashita Ashok, Barbara Bruno, Tamara Helf, and Karsten Berns. 2025. "Thanks for the Practice!": LLM-Powered Social Robot as Tandem Language Partner at University. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Melbourne, Australia, 1221–1226. doi:10.1109/HRI61500.2025.10973837

[14] Ashita Ashok, Lan Nguyen, and Karsten Berns. 2025. "A Glimpse Into My World": Empathy Towards Emotional Robot Backstories at University. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE Press, Melbourne, Australia, 1227–1231. doi:10.1109/HRI61500.2025.10973995

[15] Ruth Aylett, Michael Kriegel, Iain Wallace, Elena Márquez Segura, Johanna Mecurio, Stina Nylander, and Patricia Vargas. 2013. Do I Remember You? Memory and Identity in Multiple Embodiments. In *2013 IEEE RO-MAN*. IEEE, Gyeongju, South Korea, 143–148. doi:10.1109/ROMAN.2013.6628435

[16] Alexandra Bacula, Kamron Kayhani, Jennifer McCloskey, Dana Reason, and Heather Knight. 2020. Dance prototyping: communicating group membership and relational attitudes via multi-robot expressive motion. In *Companion of the 2020 Robotics Science and Systems Conference*. Virtual, 4 pages. https://www.charismarobotics.com/s/2020_Workshop_Dance_Prototyping_Bacula.pdf

[17] Karen Barad. 1999. Agential Realism: Feminist Interventions in Understanding Scientific Practices. In *The Science Studies Reader*, Mario Biagioli (Ed.). Routledge, 1–11.

[18] Xabier E. Barandiaran, Ezequiel Di Paolo, and Marieke Rohde. 2009. Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action. *Adaptive Behavior* 17, 5 (Oct. 2009), 367–386. doi:10.1177/1059712309343819

[19] Christoph Bartneck, Takayuki Kanda, Omar Mubin, and Abdullah Al Mahmud. 2009. Does the Design of a Robot Influence Its Animacy and Perceived Intelligence? *International Journal of Social Robotics* 1, 2 (April 2009), 195–204. doi:10.1007/s12369-009-0013-7

[20] Alexandra Bejarano, Samantha Reig, Priyanka Senapati, and Tom Williams. 2022. You Had Me at Hello: The Impact of Robot Group Presentation Strategies on Mental Model Formation. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI '22)*. IEEE Press, Sapporo, Hokkaido, Japan, 363–371. doi:10.1109/HRI53351.2022.9889465

[21] Alexandra Bejarano and Tom Williams. 2022. Understanding and Influencing User Mental Models of Robot Identity. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI '22)*. IEEE Press, Sapporo, Hokkaido, Japan, 1149–1151. doi:10.1109/HRI53351.2022.9889473

[22] Alexandra Bejarano and Tom Williams. 2023. No Name, No Voice, Less Trust: Robot Group Identity Performance, Entitativity, and Trust Distribution. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Busan, Korea, Republic of, 1339–1346. doi:10.1109/RO-MAN57019.2023.10309416

[23] Pier Paolo Bellini. 2015. Time and Social Identity. *Italian Sociological Review* 5, 1 (April 2015), 63–63. doi:10.13136/isr.v5i1.95

[24] Jasmin Bernotat, Friederike Eyssel, and Janik Sachse. 2017. Shape It – The Influence of Robot Body Shape on Gender Perception in Robots. In *Social Robotics*, Abderrahmane Kheddar, Eiichi Yoshida, Shuzhi Sam Ge, Kenji Suzuki, John-John Cabibihan, Friederike Eyssel, and Hongsheng He (Eds.). Springer International Publishing, Cham, 75–84. doi:10.1007/978-3-319-70022-9_8

[25] Jasmin Bernotat, Friederike Eyssel, and Janik Sachse. 2021. The (Fe)male Robot: How Robot Body Shape Impacts First Impressions and Trust Towards Robots. *International Journal of Social Robotics* 13, 3 (June 2021), 477–489. doi:10.1007/s12369-019-00562-7

[26] Karsten Berns and Ashita Ashok. 2024. "You Scare Me": The Effects of Humanoid Robot Appearance, Emotion, and Interaction Skills on Uncanny Valley Phenomenon. *Actuators* 13, 10 (Oct. 2024), 419. doi:10.3390/act13100419

[27] Shreyas Bhat, Joseph B. Lyons, Cong Shi, and X. Jessie Yang. 2024. Evaluating the Impact of Personalized Value Alignment in Human-Robot Interaction: Insights into Trust and Team Performance Outcomes. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 32–41. doi:10.1145/3610977.3634921

[28] Annabel Blake, Marcus Carter, and Eduardo Velloso. 2026. Restoration, Exploration and Transformation: How Youth Engage Character.AI Chatbots for Feels, Fun and Finding themselves. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3772318.3790508 In press.

[29] Marah Blaurock, Martina Čaić, Mehmet Okan, and Alexander P. Henkel. 2022. Robotic role theory: an integrative review of human–robot service interaction to advance role theory in the age of social robots. *Journal of Service Management* 33, 6 (June 2022), 27–49. doi:10.1108/JOSM-09-2021-0345

[30] Margaret Boden, Joanna Bryson, Darwin Caldwell, Kerstin Dautenhahn, Lilian Edwards, Sarah Kember, Paul Newman, Vivienne Parry, Geoff Pegman, Tom Rodden, Tom Sorrell, Mick Wallis, Blay Whitby, and Alan Winfield. 2017. Principles of robotics: regulating robots in the real world. *Connection Science* 29, 2 (April 2017), 124–129. doi:10.1080/09540091.2016.1271400

[31] Karla Bransky and Penny Sweetser. 2025. Exploring the Effects of (Re)Embodiment on Perceptions of Robot Teammates in Virtual Reality Environments. In *34th IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, Eindhoven, Netherlands, 500–507. doi:10.1109/RO-MAN63969.2025.11217829

[32] Karla Bransky, Penny Sweetser, Sabrina Caldwell, and Kingsley Fletcher. 2024. Mind-Body-Identity: A Scoping Review of Multi-Embodiment. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. ACM, New York, NY, USA, Boulder, CO, USA, 11 pages. doi:10.1145/3610977.3634922

[33] Karla Bransky, Penny Sweetser, Sabrina Caldwell, and Tom Gedeon. 2026. Identity, Form, and Function: Exploring Re-Embodiment for Human-Robot Teaming in Virtual Reality Environments. *ACM Transactions on Human-Robot Interaction* (2026), 32 pages. doi:10.1145/3806398 Just Accepted.

[34] Harry Brignull. 2010. darkpatterns.org. https://old.deceptive.design/main_page/index.html

[35] Harry Brignull. 2023. *Deceptive Patterns: Exposing the Tricks Tech Companies Use to Control You*. Testimonium.

[36] Kimberly A. Brink and Henry M. Wellman. 2020. Robot teachers for children? Young children trust robots depending on their perceived accuracy and agency. *Developmental Psychology* 56, 7 (2020), 1268–1277. doi:10.1037/dev0000884

[37] Charlie Brooker and Owen Harris. 2013. Be Right Back. Television series episode. https://www.imdb.com/title/tt2290780/ Episode 1, Season 2 of *Black Mirror*, Zeppotron/Channel 4.

[38] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Vol. 81. PMLR, New York, NY, USA, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[39] Bengisu Cagiltay and Bilge Mutlu. 2024. Toward Family-Robot Interactions: A Family-Centered Framework in HRI. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 76–85. doi:10.1145/3610977.3634976

[40] Sabrina Caldwell, Penny Sweetser, Nicholas O'Donnell, Matthew J. Knight, Matthew Aitchison, Tom Gedeon, Daniel Johnson, Margot Brereton, Marcus Gallagher, and David Conroy. 2022. An Agile New Research Framework for Hybrid Human-AI Teaming: Trust, Transparency, and Transferability. *ACM Transactions on Interactive Intelligent Systems* 12, 3 (July 2022), 17:1–17:36. doi:10.1145/3514257

[41] Donald T. Campbell. 1958. Common Fate, Similarity, and Other Indices of the Status of Aggregates of Persons as Social Entities. *Behavioral Science* 3, 1 (1958), 14–25. doi:10.1002/bs.3830030103

[42] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K. Wong. 2024. DreamAvatar: Text-and-Shape Guided 3D Human Avatar Generation via Diffusion Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 958–968. doi:10.1109/CVPR52733.2024.00097

[43] Jung Ju Choi and Sonya S. Kwak. 2017. Who is this?: Identity and Presence in Robot-Mediated Communication. *Cognitive Systems Research* 43 (June 2017), 174–189. doi:10.1016/j.cogsys.2016.07.006

[44] Herbert H. Clark and Kerstin Fischer. 2023. Social robots as depictions of social agents. *Behavioral and Brain Sciences* 46 (Jan. 2023), e21. doi:10.1017/S0140525X22000668

[45] Betty Rose Connell, Mike Jones, Ron Mace, Jim Mueller, Abir Mullick, Elaine Ostroff, Jon Sanford, Ed Steinfeld, Molly Story, and Greg Vanderheiden. 1997. The Principles of Universal Design. https://web.stanford.edu/class/engr110/2007/PUD.pdf

[46] Filipa Correia, Sofia Petisca, Patrícia Alves-Oliveira, Tiago Ribeiro, Francisco S. Melo, and Ana Paiva. 2019. "I Choose... YOU!" Membership preferences in human–robot teams. *Autonomous Robots* 43, 2 (Feb. 2019), 359–373. doi:10.1007/s10514-018-9767-9

[47] Pedro Cuba. 2010. Agent Migration between Bodies and Platforms.

[48] Stephen Darwall. 2004. Respect and the Second-Person Standpoint. *Proceedings and Addresses of the American Philosophical Association* 78, 2 (2004), 43–59. doi:10.2307/3219724

[49] Jenny L. Davis. 2020. *How Artifacts Afford: The Power and Politics of Everyday Things.* MIT Press.

[50] Jenny L Davis. 2023. 'Affordances' for Machine Learning. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23).* Association for Computing Machinery, New York, NY, USA, 324–332. doi:10.1145/3593013.3594000

[51] Anna Deichler, Siyang Wang, Simon Alexanderson, and Jonas Beskow. 2023. Learning to generate pointing gestures in situated embodied conversational agents. *Frontiers in Robotics and AI* 10 (2023), 1110534. doi:10.3389/frobt.2023.1110534

[52] Nathaniel Dennler, Mina Kian, Stefanos Nikolaidis, and Maja Matarić. 2025. Designing Robot Identity: The Role of Voice, Clothing, and Task on Robot Gender Perception. *International Journal of Social Robotics* 17, 4 (April 2025), 707–728. doi:10.1007/s12369-025-01209-6

[53] Smit Desai, Mateusz Dubiel, Nima Zargham, Thomas Mildner, and Laura Spillner. 2025. Personas Evolved: Designing Ethical LLM-Based Conversational Agent Personalities. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25).* Association for Computing Machinery, New York, NY, USA, 1–4. doi:10.1145/3719160.3728624

[54] Blanca Deusdad. 2024. Ethical implications in using robots among older adults living with dementia. *Frontiers in Psychiatry* 15 (Sept. 2024), 13 pages. doi:10.3389/fpsyt.2024.1436273

[55] Pierre Dewitte. 2024. Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. *Computer Law & Security Review* 54 (Sept. 2024), 106019. doi:10.1016/j.clsr.2024.106019

[56] Suh-Yeon Dong, Bo-Kyeong Kim, Kyeongho Lee, and Soo-Young Lee. 2015. A Preliminary Study on Human Trust Measurements by EEG for Human-Machine Interactions. In *Proceedings of the 3rd International Conference on Human-Agent Interaction (HAI '15).* Association for Computing Machinery, New York, NY, USA, 265–268. doi:10.1145/2814940.2814993

[57] Kanghui Du, Dražen Brščić, Yuyi Liu, and Takayuki Kanda. 2024. Can't You See I Am Bothered? Human-inspired Suggestive Avoidance for Robots. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24).* Association for Computing Machinery, New York, NY, USA, 184–193. doi:10.1145/3610977.3634954

[58] Brian R. Duffy and Karolina Zawieska. 2012. Suspension of disbelief in social robotics. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication.* IEEE, Paris, France, 484–489. doi:10.1109/ROMAN.2012.6343798

[59] Anthony Dunne and Fiona Raby. 2024. *Speculative Everything, With a new preface by the authors: Design, Fiction, and Social Dreaming.* MIT press.

[60] Anthony Dunne and Fiona Raby. 2025. *Not Here, Not Now: Speculative Thought, Impossibility, and the Design Imagination.* MIT Press, Cambridge, MA, US.

[61] Belinda J. Dunstan and Guy Hoffman. 2023. Social Robot Morphology: Cultural Histories of Robot Design. In *Cultural Robotics: Social Robots and Their Emergent Cultural Ecologies.* Springer, Cham, 13–34. doi:10.1007/978-3-031-28138-9_2

[62] Alexis Elder. 2019. Conversation from Beyond the Grave? A Neo-Confucian Ethics of Chatbots of the Dead. *Journal of Applied Philosophy* 37, 1 (May 2019), 73–88. doi:10.1111/japp.12369

[63] Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological Review* 114, 4 (Oct. 2007), 864–886. doi:10.1037/0033-295X.114.4.864

[64] Raffaella Esposito, Alessandra Rossi, and Silvia Rossi. 2025. Deception in HRI and Its Implications: A Systematic Review. *ACM Transactions on Human-Robot Interaction* 14, 3 (April 2025), 47:1–47:26. doi:10.1145/3721297

[65] Friederike Eyssel, Dieta Kuchenbrandt, Simon Bobinger, Laura de Ruiter, and Frank Hegel. 2012. 'If you sound like me, you must be more human': on the interplay of robot and user features on human-robot acceptance and anthropomorphism. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction (HRI '12).* Association for Computing Machinery, New York, NY, USA, 125–126. doi:10.1145/2157689.2157717

[66] Luciano Floridi. 2008. The Method of Levels of Abstraction. *Minds and Machines* 18, 3 (Sept. 2008), 303–329. doi:10.1007/s11023-008-9113-7

[67] Luciano Floridi. 2025. AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis. *Philosophy & Technology* 38, 1 (March 2025), 1–27. doi:10.1007/s13347-025-00858-9

[68] Luciano Floridi and Josh Cowls. 2022. A Unified Framework of Five Principles for AI in Society. In *Machine Learning and the City.* John Wiley & Sons, Ltd, 535–545. doi:10.1002/9781119815075.ch45

[69] Luciano Floridi and J.W. Sanders. 2004. On the Morality of Artificial Agents. *Minds and Machines* 14, 3 (Aug. 2004), 349–379. doi:10.1023/B:MIND.0000035461.63578.9d

[70] Marlena R. Fraune, Yusaku Nishiwaki, Selma Sabanović, Eliot R. Smith, and Michio Okada. 2017. Threatening Flocks and Mindful Snowflakes: How Group Entitativity Affects Perceptions of Robots. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17).* Association for Computing Machinery, New York, NY, USA, 205–213. doi:10.1145/2909824.3020248

[71] Marlena R. Fraune, Steven Sherrin, Selma Sabanović, and Eliot R. Smith. 2015. Rabble of Robots Effects: Number and Type of Robots Modulates Attitudes, Emotions, and Stereotypes. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15).* Association for Computing Machinery, New York, NY, USA, 109–116. doi:10.1145/2696454.2696483

[72] Marlena R. Fraune, Selma Šabanović, and Eliot R. Smith. 2017. Teammates first: Favoring ingroup robots over outgroup humans. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* IEEE, Lisbon, Portugal, 1432–1437. doi:10.1109/ROMAN.2017.8172492

[73] Julian De Freitas, Noah Castelo, Ahmet K. Uğuralp, and Zeliha Oğuz-Uğuralp. 2025. Lessons From an App Update at Replika AI: Identity Discontinuity in Human-AI Relationships. doi:10.48550/arXiv.2412.14190

[74] Natalie Friedman, Alexandra Bremers, Bolor Amgalan, RAY LC, A. J. Parry, Kari Love, and Wendy Ju. 2024. Clothing for Robot Identity. In *Proceedings of the Robo-Identity: Designing for Identity in the Shared World Workshop, held at the 19th ACM/IEEE International Conference on Human-Robot Interaction (HRI '24).* Boulder, CO, USA, 3 pages.

[75] Natalie Friedman, Kari Love, RAY LC, Jenny E Sabin, Guy Hoffman, and Wendy Ju. 2021. What Robots Need From Clothing. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference (DIS '21).* Association for Computing Machinery, New York, NY, USA, 1345–1355. doi:10.1145/3461778.3462045

[76] Katherine K. Fu, Maria C. Yang, and Kristin L. Wood. 2016. Design Principles: Literature Review, Analysis, and Future Directions. *Journal of Mechanical Design* 138, 10:101103 (30 Aug. 2016), 13 pages. doi:10.1115/1.4034105

[77] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2025. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984v3* (2025). doi:10.48550/arXiv.2307.14984

[78] Paulo Fontaínha Gomes, Alberto Sardinha, Elena Márquez Segura, Henriette Cramer, and Ana Paiva. 2014. Migration Between Two Embodiments of an Artificial Pet. *International Journal of Humanoid Robotics* 11, 1 (2014), 1–32. doi:10.1142/S0219843614500017

[79] Paulo Fontaínha Gomes, Elena Márquez Segura, Henriette Cramer, Ana Paiva, and Lars Erik Holmquist. 2011. ViPleo and PhyPleo: Artificial Pet with Two Embodiments. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology.* Association for Computing Machinery, Lisbon, Portugal, 1–8. doi:10.1145/2071423.2071427

[80] Juhye Ha, Hyeon Jeon, Daeun Han, Jinwook Seo, and Changhoon Oh. 2024. CloChat: Understanding How People Customize, Interact, and Experience Personas in Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24).* Association for Computing Machinery, New York, NY, USA, 1–24. doi:10.1145/3613904.3642472

[81] Frank Hegel, Claudia Muhl, Britta Wrede, Martina Hielscher-Fastabend, and Gerhard Sagerer. 2009. Understanding Social Robots. In *2009 Second International Conferences on Advances in Computer-Human Interactions.* IEEE, Cancun, Mexico, 169–174. doi:10.1109/ACHI.2009.51

[82] Anna Henschel, Guy Laban, and Emily S Cross. 2021. What Makes a Robot Social? A Review of Social Robots from Science Fiction to a Home or Hospital Near You. *Current Robotics Reports* 2 (2021), 9–19. doi:10.1007/s43154-020-00035-0

[83] Damith Herath, Janie Busby Grant, Adrian Rodriguez, and Jenny L. Davis. 2025. First impressions of a humanoid social robot with natural language capabilities. *Scientific Reports* 15, 1 (June 2025), 19715. doi:10.1038/s41598-025-04274-z

[84] AI HLEG. 2019. *Ethics Guidelines for Trustworthy AI.* Technical Report. High-Level Expert Group on Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [Accessed 06-04-2026].

[85] Jerlyn Q. H. Ho, Meilan Hu, Tracy X. Chen, and Andree Hartanto. 2025. Potential and pitfalls of romantic Artificial Intelligence (AI) companions: A systematic review. *Computers in Human Behavior Reports* 19 (Aug. 2025), 100715. doi:10.1016/j.chbr.2025.100715

[86] Thomas Hobbes. 2021. *Elements of Philosophy, the First Section, Concerning Body.* Legare Street Press. https://philpapers.org/rec/HOBEOP

[87] Guy Hoffman. 2011. On stage: robots as performers. In *Proceedings of the RSS 2011 workshop on human-robot interaction: Perspectives and contributions to robotics from the human sciences (2011).* Los Angeles, CA, 5 pages. https://hrc2.io/assets/pdfs/papers/HoffmanWkRSS11.pdf

[88] Colin Holbrook, Daniel Holman, Joshua Clingo, and Alan R. Wagner. 2024. Overtrust in AI Recommendations About Whether or Not to Kill: Evidence from Two Human-Robot Interaction Studies. *Scientific Reports* 14, 1 (Sept. 2024),

19751. doi:10.1038/s41598-024-69771-z

[89] Patrick Holthaus, Catherine Menon, and Farshid Amirabdollahian. 2019. How a Robot's Social Credibility Affects Safety Performance. In *International Conference on Social Robotics (ICSR 2019)*, Miguel A. Salichs, Shuzhi Sam Ge, Emilia Ivanova Barakova, John-John Cabibihan, Alan R. Wagner, Álvaro Castro-González, and Hongsheng He (Eds.). Lecture Notes in Computer Science, Vol. 11876. Springer Cham, Madrid, Spain, 740–749. doi:10.1007/978-3-030-35888-4_69

[90] Patrick Holthaus, Trenton Schulz, Gabriella Lakatos, and Rebekka Soma. 2023. Communicative Robot Signals: Presenting a New Typology for Human-Robot Interaction. In *HRI '23: Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. ACM/IEEE, Stockholm, Sweden, 132–141. doi:10.1145/3568162.3578631

[91] Patrick Holthaus and Sven Wachsmuth. 2021. It was a Pleasure Meeting You - Towards a Holistic Model of Human-Robot Encounters. *International Journal of Social Robotics (Springer)* 13, 7 (2021), 1729–1745. doi:10.1007/s12369-021-00759-9

[92] Ying-yi Hong, Siran Zhan, Michael W Morris, and Verónica Benet-Martínez. 2016. Multicultural identity processes. *Current Opinion in Psychology* 8 (2016), 49–53. doi:10.1016/j.copsyc.2015.09.020

[93] Matthew J Hornsey. 2008. Social identity theory and self-categorization theory: A historical review. *Social and personality psychology compass* 2, 1 (2008), 204–222. doi:10.1111/j.1751-9004.2007.00066.x

[94] Aike C. Horstmann and Nicole C. Krämer. 2019. Great Expectations? Relation of Previous Experiences With Social Robots in Real Life or in the Media and Expectancies Based on Qualitative and Quantitative Assessment. *Frontiers in Psychology* 10 (30 Apr 2019), 939. doi:10.3389/fpsyg.2019.00939

[95] Yoyo Tsung-Yu Hou, EunJeong Cheon, and Malte F. Jung. 2024. Power in Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 269–282. doi:10.1145/3610977.3634949

[96] Yi Hu, Shijia Kang, Haotong Yang, Haotian Xu, and Muhan Zhang. 2025. Training Large Language Models to be Better Rule Followers. doi:10.48550/arXiv.2502.11525

[97] Guanyu Huang and Roger K. Moore. 2022. Is honesty the best policy for mismatched partners? Aligning multi-modal affordances of a social robot: An opinion paper. *Frontiers in Virtual Reality* 3, 1020169 (16 Sept. 2022), 5 pages. doi:10.3389/frvir.2022.1020169

[98] Guanyu Huang and Roger K. Moore. 2025. Adaptive Affordance Design for Social Robots: Tailoring to Role-Specific Preferences. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Melbourne, Australia, 580–588. doi:10.1109/HRI61500.2025.10974246

[99] Ohad Inbar and Noam Tractinsky. 2009. The Incidental User. *Interactions* 16, 4 (Jan. 2009), 56–59. https://doi.org/10.1145/1551986.1551998

[100] Ryan Blake Jackson and Tom Williams. 2021. A Theory of Social Agency for Human-Robot Interaction. *Frontiers in Robotics and AI* 8 (13 Aug. 2021), 1–15. doi:10.3389/frobt.2021.687726

[101] Rucha Khot, Teis Arets, Joel Wester, Franziska Burger, Niels Van Berkel, Rens Brankaert, Wijnand IJsselsteijn, and Minha Lee. 2025. Challenging Futures: Using Chatbots to Reflect on Aging and Dementia. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. doi:10.1145/3706598.3713727

[102] Rucha Khot, Minha Lee, Alexandra Bejarano, Lux Miranda, Gisela Reyes-Cruz, Joel E. Fischer, and Dimosthenis Kontogiorgos. 2024. Robo-Identity: Designing for Identity in the Shared World. In *HRI '24: Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Boulder, CO, USA, 1326–1328. doi:10.1145/3610978.3638166

[103] Tobias Kirschstein, Simon Giebenhain, and Matthias Nießner. 2024. DiffusionAvatars: Deferred Diffusion for High-fidelity 3D Head Avatars. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 5481–5492. doi:10.1109/CVPR52733.2024.00524

[104] René F. Kizilcec. 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2390–2395. doi:10.1145/2858036.2858402

[105] Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. 2024. Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior* 160 (Nov. 2024), 108352. doi:10.1016/j.chb.2024.108352

[106] Heather Knight. 2011. Eight Lessons Learned about Non-verbal Interactions through Robot Theater. In *Social Robotics*, Bilge Mutlu, Christoph Bartneck, Jaap Ham, Vanessa Evers, and Takayuki Kanda (Eds.). Vol. 7072. Springer Berlin Heidelberg, Berlin, Heidelberg, 42–51. doi:10.1007/978-3-642-25504-5_5

[107] Kheng Lee Koay, Dag Sverre Syrdal, Kerstin Dautenhahn, Krzysztof Arent, Łukasz Małek, and Bogdan Kreczmer. 2011. Companion Migration – Initial Participants' Feedback from a Video-Based Prototyping Study. In *Mixed Reality and Human-Robot Interaction*, Xiangyu Wang (Ed.). Springer Netherlands, Dordrecht, 133–151. doi:10.1007/978-94-007-0582-1_8

[108] Kheng Lee Koay, Dag Sverre Syrdal, Wan Ching Ho, and Kerstin Dautenhahn. 2016. Prototyping Realistic Long-Term Human-Robot Interaction for the Study of Agent Migration. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE Press, New York, NY, USA, 809–816. doi:10.1109/ROMAN.2016.7745212

[109] Michael Kriegel, Ruth Aylett, Pedro Cuba, Marco Vala, and Ana Paiva. 2011. Robots Meet IVAs: A Mind-Body Interface for Migrating Artificial Intelligent Agents. In *Intelligent Virtual Agents (Lecture Notes in Computer Science)*, Hannes Högni Vilhjálmsson, Stefan Kopp, Stacy Marsella, and Kristinn R. Thórisson (Eds.). Springer, Berlin, Heidelberg, 282–295. doi:10.1007/978-3-642-23974-8_31

[110] Joel Krueger and Lucy Osler. 2022. Communing with the Dead Online: Chatbots, Grief, and Continuing Bonds. *Journal of Consciousness Studies* 29, 9-10 (Sept. 2022), 222–252. doi:10.53765/20512201.29.9.222

[111] Minae Kwon, Malte F. Jung, and Ross A. Knepper. 2016. Human expectations of social robots. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Christchurch, New Zealand, 463–464. doi:10.1109/HRI.2016.7451807

[112] Guy Laban, Sophie Chiang, and Hatice Gunes. 2025. What People Share With a Robot When Feeling Lonely and Stressed and How It Helps Over Time. In *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Eindhoven, Netherlands, 1930–1935. doi:10.1109/RO-MAN63969.2025.11217573

[113] Guy Laban and Emily. S. Cross. 2024. Sharing our Emotions with Robots: Why do we do it and how does it make us feel? *IEEE Transactions on Affective Computing (Early Access)* (2024), 1–18. doi:10.1109/TAFFC.2024.3470984

[114] Guy Laban, Jean-Noël George, Val Morrison, and Emily S. Cross. 2021. Tell me more! Assessing interactions with social robots from speech. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 136–159. doi:10.1515/pjbr-2021-0011

[115] Guy Laban, Arvid Kappas, Val Morrison, and Emily S. Cross. 2024. Building Long-Term Human–Robot Relationships:Examining Disclosure, Perception and Well-Being Across Time. *International Journal of Social Robotics* 16, 5 (May 2024), 1–27. doi:10.1007/s12369-023-01076-z

[116] Guy Laban, Sébastien Le Maguer, Minha Lee, Dimosthenis Kontogiorgos, Samantha Reig, Ilaria Torre, Ravi Tejwani, Matthew J. Dennis, and Andre Pereira. 2022. Robo-Identity: Exploring Artificial Identity and Emotion via Speech Interactions. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Sapporo, Japan, 1265–1268. doi:10.1109/HRI53351.2022.9889649

[117] Guy Laban, Val Morrison, Arvid Kappas, and Emily S. Cross. 2025. Coping with Emotional Distress via Self-Disclosure to Robots: An Intervention with Caregivers. *International Journal of Social Robotics* 17, 9 (Sept. 2025), 1837–1870. doi:10.1007/s12369-024-01207-0

[118] Guy Laban, Julie Wang, and Hatice Gunes. 2026. A Robot-Led Intervention for Emotion Regulation: from Expression to Reappraisal. *IEEE Transactions on Affective Computing (Early Access)* (26 Jan. 2026), 1–15. doi:10.1109/TAFFC.2026.3657604

[119] Weston Laity, Patrick Holthaus, and Kerstin Haring. 2025. Robot Continuity across Embodiments: Portability, Identity and Migration of Robotic Systems. In *34th IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, Eindhoven, Netherlands, 805–812. doi:10.1109/RO-MAN63969.2025.11217771

[120] Amy LaViers, Catie Cuan, Catherine Maguire, Karen Bradley, Kim Brooks Mata, Alexandra Nilles, Ilya Vidrin, Novoneel Chakraborty, Madison Heimerdinger, and Umer Huzaifa. 2018. Choreographic and somatic approaches for the development of expressive robotic systems. *Arts* 7, 2 (March 2018), 11. doi:10.3390/arts7020011

[121] Steven Lawrence, Melanie Jouaiti, Jesse Hoey, Chrystopher L. Nehaniv, and Kerstin Dautenhahn. 2025. The Role of Social Norms in Human–Robot Interaction: A Systematic Review. *ACM Transactions on Human-Robot Interaction* 14, 3 (June 2025), 56:1–56:44. doi:10.1145/3722120

[122] Kiljae Lee, Kyung Young Lee, and Lorn Sheehan. 2020. Hey Alexa! A Magic Spell of Social Glue?: Sharing a Smart Voice Assistant Speaker and Its Impact on Users' Perception of Group Harmony. *Information Systems Frontiers* 22, 3 (June 2020), 563–583. doi:10.1007/s10796-019-09975-1

[123] Minha Lee, Dimosthenis Kontogiorgos, Ilaria Torre, Michal Luria, Ravi Tejwani, Matthew J. Dennis, and Andre Pereira. 2021. Robo-Identity: Exploring Artificial Identity and Multi-Embodiment. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion)*. ACM, New York, NY, USA, 718–720. doi:10.1145/3434074.3444878

[124] Minha Lee, Gale Lucas, Johnathan Mell, Emmanuel Johnson, and Jonathan Gratch. 2019. What's on Your Virtual Mind? Mind Perception in Human-Agent Negotiations. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (IVA '19)*. Association for Computing Machinery, New York, NY, USA, 38–45. doi:10.1145/3308532.3329465

[125] Minha Lee, Peter Ruijten, Lily Frank, Yvonne de Kort, and Wijnand IJsselsteijn. 2021. People May Punish, But Not Blame Robots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3411764.3445284

[126] Minha Lee, Peter Ruijten, Lily Frank, and Wijnand IJsselsteijn. 2023. Here's Looking at You, Robot: The Transparency Conundrum in HRI. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Busan, Korea, Republic of, 2120–2127. doi:10.1109/RO-MAN57019.2023.10309653

[127] Hans-Thies Lehmann and Karen Jürs-Munby. 2006. *Postdramatic theatre* (1st ed.). Routledge, London. 224 pages. doi:10.4324/9780203088104

[128] Benedikt Leichtmann, Verena Nitsch, and Martina Mara. 2022. Crisis Ahead? Why Human-Robot Interaction User Studies May Have Replicability Problems and Directions for Improvement. *Frontiers in Robotics and AI* 9 (2022), 1–15. doi:10.3389/frobt.2022.838116

[129] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems* 36 (2023), 51991–52008. https://proceedings.neurips.cc/paper_files/paper/2023/file/a3621ee907def47c1b952ade25c67698-Paper-Conference.pdf

[130] Shimin Li, Tianxiang Sun, Qinyuan Cheng, and Xipeng Qiu. 2024. Agent Alignment in Evolving Social Norms. doi:10.48550/arXiv.2401.04620

[131] Brian Lickel, David L. Hamilton, Grazyna Wieczorkowska, Amy Lewis, Steven J. Sherman, and A. Neville Uhles. 2000. Varieties of groups and the perception of group entitativity. *Journal of Personality and Social Psychology* 78, 2 (2000), 223–246. doi:10.1037/0022-3514.78.2.223

[132] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. Association for Computing Machinery, New York, NY, USA, 2119–2128. doi:10.1145/1518701.1519023

[133] Dongshuo Liu, Zhijing Wu, Dandan Song, and Heyan Huang. 2025. A Persona-Aware LLM-Enhanced Framework for Multi-Session Personalized Dialogue Generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 103–123. doi:10.18653/v1/2025.findings-acl.5

[134] Maria Luce Lupetti, Cristina Zaga, and Nazli Cila. 2021. Designerly Ways of Knowing in HRI: Broadening the Scope of Design-oriented HRI Through the Concept of Intermediate-level Knowledge. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21)*. Association for Computing Machinery, New York, NY, USA, 389–398. doi:10.1145/3434073.3444668

[135] Michal Luria, Samantha Reig, Xiang Zhi Tan, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2019. Re-Embodiment and Co-Embodiment: Exploration of Social Presence for Robots and Conversational Agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19)*. Association for Computing Machinery, New York, NY, USA, 633–644. doi:10.1145/3322276.3322340

[136] Joseph B. Lyons, Garrett G. Sadler, Kolina Koltai, Henri Battiste, Nhut T. Ho, Lauren C. Hoffmann, David Smith, Walter Johnson, and Robert Shively. 2017. Shaping Trust Through Transparent Design: Theoretical and Experimental Guidelines. In *Advances in Human Factors in Robots and Unmanned Systems*, Pamela Savage-Knepshield and Jessie Chen (Eds.). Springer International Publishing, Cham, 127–136. doi:10.1007/978-3-319-41959-6_11

[137] Jacob P. Macdonald, Rohit Mallick, Allan B. Wollaber, Jaime D. Peña, Nathan McNeese, and Ho Chit Siu. 2024. Language, Camera, Autonomy! Prompt-engineered Robot Control for Rapidly Evolving Deployment. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 717–721. doi:10.1145/3610978.3640671

[138] Alan Martin, Gregory M. P. O'Hare, Brian R. Duffy, Bianca Schön, and John F. Bradley. 2005. Maintaining the Identity of Dynamically Embodied Agents. In *Intelligent Virtual Agents (Lecture Notes in Computer Science)*, Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier, and Thomas Rist (Eds.). Springer, Berlin, Heidelberg, 454–465. doi:10.1007/11550617_38

[139] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. 2025. *The AI Index 2025 Annual Report.* Technical Report. Institute for Human-Centered AI, Stanford University, Stanford, CA. 456 pages. doi:10.48550/arXiv.2504.07139

[140] Lux Miranda. 2024. Who is a robot? A fundamental model of artificial identity. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, Pasadena, CA, USA, 1510–1515. doi:10.1109/RO-MAN60168.2024.10731452

[141] Lux Miranda, Ginevra Castellano, and Katie Winkle. 2023. Examining the State of Robot Identity. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23)*. Association for Computing Machinery, New York, NY, USA, 658–662. doi:10.1145/3568294.3580168

[142] Lux Miranda, Ginevra Castellano, and Katie Winkle. 2024. A Case for Diverse Social Robot Identity Performance in Education. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Boulder, CO, USA, 28–35. doi:10.1145/3610978.3640768

[143] Byeong June Moon, Jong Suk Choi, and Sonya S. Kwak. 2021. Pretending to be Okay in a Sad Voice: Social Robot's Usage of Verbal and Nonverbal Cue Combination and its Effect on Human Empathy and Behavior Inducement. In *IEEE International Conference on Intelligent Robots and Systems*. IEEE, Prague, Czech Republic, 854–861. doi:10.1109/IROS51168.2021.9636709

[144] Youngme Moon and Clifford Nass. 1996. How "Real" Are Computer Personalities?: Psychological Responses to Personality Types in Human-Computer Interaction. *Communication Research* 23, 6 (Dec. 1996), 651–674. doi:10.1177/009365096023006002

[145] Roger K. Moore. 2012. A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. *Scientific Reports* 2, 1 (Nov. 2012), 864. doi:10.1038/srep00864

[146] Roger K. Moore. 2016. Introducing a Pictographic Language for Envisioning a Rich Variety of Enactive Systems with Different Degrees of Complexity. *International Journal of Advanced Robotic Systems* 13, 2 (March 2016), 74. doi:10.5772/62244

[147] Roger K. Moore. 2019. A 'Canny' Approach to Spoken Language Interfaces. In *CHI-19 Workshop on Mapping Theoretical and Methodological Perspectives for Understanding Speech Interface Interactions*. ACM, Glasgow, Scotland UK, 3 pages. doi:10.48550/arXiv.1908.08131

[148] Mashiro Mori. 1970. Bukimi no tani (the uncanny valley). *Energy* 7 (1970), 33–35.

[149] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics \& Automation Magazine* 19, 2 (June 2012), 98–100. doi:10.1109/MRA.2012.2192811

[150] Fabian Muhly, Emanuele Chizzonic, and Philipp Leo. 2025. AI-deepfake scams and the importance of a holistic communication security strategy. *International Cybersecurity Law Review* 6, 1 (March 2025), 53–61. doi:10.1365/s43439-025-00143-7

[151] Bonnie M. Muir. 1987. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies* 27, 5 (Nov. 1987), 527–539. doi:10.1016/S0020-7373(87)80013-5

[152] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (2000), 81–103. doi:10.1111/0022-4537.00153

[153] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. Association for Computing Machinery, Boston, Massachusetts, USA, 72–78. doi:10.1145/191666.191703

[154] Alan F. Newell, Margaret E. Morgan, Lorna Gibson, and Paula Forbes. 2011. Experiences with professional theatre for awareness raising. *Interacting with Computers* 23, 6 (Nov. 2011), 594–603. doi:10.1016/j.intcom.2011.08.002

[155] Massimiliano Nigro, Emmanuel Akinrintoyo, Nicole Salomons, and Micol Spitale. 2025. Social Group Human-Robot Interaction: A Scoping Review of Computational Challenges. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI '25')*. IEEE, Melbourne, Australia, 468–478. doi:10.1109/HRI61500.2025.10973980

[156] Don Norman. 2013. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books, New York, NY, USA.

[157] Ministry of Culture Denmark. 2025. Bred aftale om deepfakes giver alle ret til egen krop og egen stemme. https://kum.dk/aktuelt/nyheder/bred-aftale-om-deepfakes-giver-alle-ret-til-egen-krop-og-egen-stemme [Accessed 06-04-2026].

[158] Kohei Ogawa and Tetsuo Ono. 2008. ITACO: Effects to Interactions by Relationships between Humans and Artifacts. In *Intelligent Virtual Agents: 8th International Conference*. Springer, Tokyo, Japan, 296–307. doi:10.1007/978-3-540-85483-8_31

[159] Kohei Okuoka, Kouichi Enami, Mitsuhiko Kimoto, and Michita Imai. 2022. Multi-Device Trust Transfer: Can Trust Be Transferred Among Multiple Devices? *Frontiers in Psychology* 13 (Aug. 2022), Article 920844. doi:10.3389/fpsyg.2022.920844

[160] Anastasia K. Ostrowski, Raechel Walker, Madhurima Das, Maria Yang, Cynthia Breazeal, Hae Won Park, and Aditi Verma. 2022. Ethics, Equity, & Justice in Human-Robot Interaction: A Review and Future Directions. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Napoli, Italy, 969–976. doi:10.1109/RO-MAN53752.2022.9900805

[161] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3586183.3606763

[162] Giulia Perugia, Stefano Guidi, Margherita Bicchi, and Oronzo Parlangeli. 2022. The Shape of Our Bias: Perceived Age and Gender in the Humanoid Robots of the ABOT Database. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Sapporo, Japan, 110–119. doi:10.1109/HRI53351.

2022.9889366
[163] Giulia Perugia and Dominika Lisy. 2023. Robot's Gendering Trouble: A Scoping Review of Gendering Humanoid Robots and Its Effects on HRI. *International Journal of Social Robotics* 15, 11 (Nov. 2023), 1725–1753. doi:10.1007/s12369-023-01061-6
[164] Torr Polakow, Guy Laban, Andrei Teodorescu, Jerome R. Busemeyer, and Goren Gordon. 2022. Social robot advisors: effects of robot judgmental fallacies and context. *Intelligent Service Robotics* 15, 5 (2022), 593–609. doi:10.1007/s11370-022-00438-2
[165] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3174214
[166] Mirjana Prpa, Giovanni Troiano, Bingsheng Yao, Toby Jia-Jun Li, Dakuo Wang, and Hansu Gu. 2024. Challenges and Opportunities of LLM-Based Synthetic Personae and Data in HCI. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing (CSCW Companion '24)*. Association for Computing Machinery, New York, NY, USA, 716–719. doi:10.1145/3678884.3681826
[167] John Pruitt and Jonathan Grudin. 2003. Personas: practice and theory. In *Proceedings of the 2003 conference on Designing for user experiences (DUX '03)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/997078.997089
[168] Luca Raggioli, Raffaella Esposito, Alessandra Rossi, and Silvia Rossi. 2025. Exploring the Role of Robot's Movements for a Transparent Affective Communication. *IEEE Robotics and Automation Letters* 10, 5 (May 2025), 4364–4371. doi:10.1109/LRA.2025.3548412
[169] Robert Ranisch and Joschka Haltaufderheide. 2025. Rapid Integration of LLMs in Healthcare Raises Ethical Concerns: An Investigation into Deceptive Patterns in Social Robots. *Digital Society* 4, 1 (Feb. 2025), 7. doi:10.1007/s44206-025-00161-2
[170] Samantha Reig, Elizabeth J. Carter, Terrence Fong, Jodi Forlizzi, and Aaron Steinfeld. 2021. Flailing, Hailing, Prevailing: Perceptions of Multi-Robot Failure Recovery Strategies. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21)*. Association for Computing Machinery, New York, NY, USA, 158–167. doi:10.1145/3434073.3444659
[171] Samantha Reig, Michal Luria, Elsa Forberger, Isabel Won, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2021. Social Robots in Service Contexts: Exploring the Rewards and Risks of Personalization and Re-embodiment. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference (DIS '21)* (Virtual Event, USA). Association for Computing Machinery, New York, NY, USA, 1390–1402. doi:10.1145/3461778.3462036
[172] Samantha Reig, Michal Luria, Janet Z. Wang, Danielle Oltman, Elizabeth Jeanne Carter, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2020. Not Some Random Agent: Multi-person Interaction with a Personalizing Service Robot. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20)*. Association for Computing Machinery, New York, NY, USA, 289–297. doi:10.1145/3319502.3374795
[173] David Robert and Cynthia Breazeal. 2012. Blended Reality Characters. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction (HRI '12)*. Association for Computing Machinery, New York, NY, USA, 359–366. doi:10.1145/2157689.2157810
[174] Amir Rosenmann, Gerhard Reese, and James E. Cameron. 2016. Social Identities in a Globalized World: Challenges and Opportunities for Collective Action. *Perspectives on Psychological Science* 11, 2 (March 2016), 202–221. doi:10.1177/1745691615621272
[175] Julia Rosén, Jessica Lindblom, and Erik Billing. 2022. The Social Robot Expectation Gap Evaluation Framework. In *Human-Computer Interaction. Technological Innovation*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 590–610. doi:10.1007/978-3-031-05409-9_43
[176] Julia Rosén, Jessica Lindblom, Maurice Lamb, and Erik Billing. 2024. Previous Experience Matters: An in-Person Investigation of Expectations in Human–Robot Interaction. *International Journal of Social Robotics* 16, 3 (March 2024), 447–460. doi:10.1007/s12369-024-01107-3
[177] Adam Rule and Jodi Forlizzi. 2012. Designing interfaces for multi-user, multi-robot systems. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction (HRI '12)*. Association for Computing Machinery, New York, NY, USA, 97–104. doi:10.1145/2157689.2157705
[178] Jonathan Rutherford. 1990. *Identity: Community, Culture, Difference*. Lawrence & Wishart, London.
[179] Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. PersonaGym: Evaluating Persona Agents and LLMs. *arXiv preprint arXiv:2407.18416* (2024), 24 pages. doi:10.48550/arXiv.2407.18416
[180] Lindsay Sanneman and Julie A. Shah. 2023. Validating metrics for reward alignment in human-autonomy teaming. *Computers in Human Behavior* 146 (Sept. 2023), 107809. doi:10.1016/j.chb.2023.107809
[181] Maggi Savin-Baden and David Burden. 2019. Digital Immortality and Virtual Humans. *Postdigital Science and Education* 1, 1 (April 2019), 87–103. doi:10.1007/s42438-018-0007-6
[182] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251. doi:10.1145/3301275.3302308
[183] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional HCI: Engaging Identity through Gender, Race, and Class. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 5412–5427. doi:10.1145/3025453.3025766
[184] Melanie Schmidt-Wolf, Tyler Becker, Denielle Oliva, Monica Nicolescu, and David Feil-Seifer. 2024. Investigating Non-Verbal Cues in Cluttered Environments: Insights Into Legible Motion From Interpersonal Interaction. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. IEEE, Pasadena, CA, USA, 1250–1257. doi:10.1109/RO-MAN60168.2024.10731388
[185] Seth J. Schwartz, Byron L. Zamboanga, and Robert S. Weisskirch. 2008. Broadening the Study of the Self: Integrating the Study of Personal Identity and Cultural Identity. *Social and Personality Psychology Compass* 2, 2 (2008), 635–651. doi:10.1111/j.1751-9004.2008.00077.x
[186] Katie Seaborn. 2022. From Identified to Self-Identifying: Social Identity Theory for Socially Embodied Artificial Agents. https://www.researchgate.net/profile/Katie-Seaborn/publication/359756008
[187] Katie Seaborn. 2025. Social Identity in Human-Agent Interaction: A Primer. *ACM Transactions on Human-Robot Interaction* (Aug. 2025). doi:10.1145/3760500 Just Accepted.
[188] Isabella Seeber, Eva Bittner, Robert O. Briggs, Triparna de Vreede, Gert-Jan de Vreede, Aaron Elkins, Ronald Maier, Alexander B. Merz, Sarah Oeste-Reiß, Nils Randrup, Gerhard Schwabe, and Matthias Söllner. 2020. Machines as teammates: A research agenda on AI in team collaboration. *Information & Management* 57, 2 (March 2020), 103174. doi:10.1016/j.im.2019.103174
[189] Elena Márquez Segura, Henriette Cramer, Paulo Fontaínha Gomes, Stina Nylander, and Ana Paiva. 2012. Revive! Reactions to Migration between Different Embodiments when Playing with Robotic Pets. In *Proceedings of the 11th International Conference on Interaction Design and Children (IDC '12)*. Association for Computing Machinery, New York, NY, USA, 88–97. doi:10.1145/2307096.2307107
[190] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, Calgary, AB, Canada, 4779–4783. doi:10.1109/ICASSP.2018.8461368
[191] Claudia Sinoo, Sylvia van der Pal, Olivier A. Blanson Henkemans, Anouk Keizer, Bert P. B. Bierman, Rosemarijn Looije, and Mark A. Neerincx. 2018. Friendship with a Robot: Children's Perception of Similarity between a Robot's Physical and Virtual Embodiment that Supports Diabetes Self-Management. *Patient Education and Counseling* 101, 7 (2018), 1248–1255. doi:10.1016/j.pec.2018.02.008
[192] Paul E. Smaldino. 2019. Social identity and cooperation in cultural evolution. *Behavioural Processes* 161 (April 2019), 108–116. doi:10.1016/j.beproc.2017.11.015
[193] Micol Spitale, Minja Axelsson, Sooyeon Jeong, Paige Tuttösí, Caitlin A. Stamatis, Guy Laban, Angelica Lim, and Hatice Gunes. 2025. Past, Present, and Future: A Survey of The Evolution of Affective Robotics For Well-being. *IEEE Transactions on Affective Computing (Early Access)* (2025), 1–17. doi:10.1109/TAFFC.2025.3567740
[194] Samantha Stedtler. 2024. Social Injustice, Group Membership and Epistemic Trust in Robots. https://sites.google.com/view/hri2024workshop-robo-identity3/accepted-papers
[195] Samantha Stedtler and Marianna Leventi. 2025. Who Is Responsible? Social Identity, Robot Errors and Blame Attribution. *Social Robots with AI: Prospects, Risks, and Responsible Methods* 397 (2025), 284–297. doi:10.3233/FAIA241515
[196] Peter Frederick Strawson. 2008 [1963]. *Freedom and Resentment and Other Essays*. Routledge, Abingdon, UK.
[197] Dag Sverre Syrdal, Kerstin Dautenhahn, Kheng Lee Koay, and Wan Ching Ho. 2015. Integrating Constrained Experiments in Long-Term Human–Robot Interaction Using Task and Scenario-Based Prototyping. *The Information Society* 31, 3 (May 2015), 265–283. doi:10.1080/01972243.2015.1020212
[198] Dag Sverre Syrdal, Kerstin Dautenhahn, Michael L. Walters, Kheng Lee Koay, and Nuno R. Otero. 2011. The Theatre methodology for facilitating discussion in human-robot interaction on information disclosure in a home environment. In *2011 RO-MAN*. IEEE, Atlanta, GA, USA, 479–484. doi:10.1109/ROMAN.2011.6005247 ISSN: 1944-9437.
[199] Dag Sverre Syrdal, Michael L. Walters, Nuno Otero, Kheng Lee Koay, and Kerstin Dautenhahn. 2007. "He Knows When You Are Sleeping": Privacy and the Personal Robot Companion. In *Proceedings of the Workshop Human Implications of Human-Robot Interaction, Association for the Advancement of Artificial Intelligence (AAAI'07)*. AAAI Press, 28–33.
[200] Martin Sökefeld. 1999. Debating Self, Identity, and Culture in Anthropology. *Current Anthropology* 40, 4 (Aug. 1999), 417–448. doi:10.1086/200042

[201] Henri Tajfel. 1982. Social psychology of intergroup relations. *Annual Review of Psychology* 33 (1982), 1–39. doi:10.1146/annurev.ps.33.020182.000245

[202] Henri Tajfel, M. G. Billig, R. P. Bundy, and Claude Flament. 1971. Social categorization and intergroup behaviour. *European Journal of Social Psychology* 1, 2 (1971), 149–178. doi:10.1002/ejsp.2420010202

[203] Ravi Tejwani, Boris Katz, and Cynthia Breazeal. 2021. Migratable AI : Investigating Users' Affect on Identity and Information Migration of a Conversational AI Agent. In *Social Robotics (Lecture Notes in Computer Science, Vol. 13086)*, Haizhou Li, Shuzhi Sam Ge, Yan Wu, Agnieszka Wykowska, Hongsheng He, Xiaorui Liu, Dongyu Li, and Jairo Perez-Osorio (Eds.). Springer International Publishing, Cham, 257–267. doi:10.1007/978-3-030-90525-5_22

[204] Ravi Tejwani, Boris Katz, and Cynthia Breazeal. 2022. Migratable AI: Personalizing Dialog Conversations with Migration Context. In *Social Robotics (Lecture Notes in Computer Science, Vol. 13817)*, Filippo Cavallo, John-John Cabibihan, Laura Fiorini, Alessandra Sorrentino, Hongsheng He, Xiaorui Liu, Yoshio Matsumoto, and Shuzhi Sam Ge (Eds.). Springer, Cham, 89–99. doi:10.1007/978-3-031-24667-8_8

[205] Ravi Tejwani, Felipe Moreno, Sooyeon Jeong, Hae Won Park, and Cynthia Breazeal. 2020. Migratable AI: Effect of Identity and Information Migration on Users' Perception of Conversational AI Agents. In *RO-MAN 2020 - 29th IEEE International Conference on Robot and Human Interactive Communication*. IEEE, Naples, Italy, 877–884. doi:10.1109/RO-MAN47096.2020.9223436

[206] Ilaria Torre and Laurence White. 2021. Trust in Vocal Human–Robot Interaction: Implications for Robot Voice Design. In *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers*, Benjamin Weiss, Jürgen Trouvain, Melissa Barkat-Defradas, and John J. Ohala (Eds.). Springer, Singapore, 299–316. doi:10.1007/978-981-15-6627-1_16

[207] Rudolph Triebel, Kai Arras, Rachid Alami, Lucas Beyer, Stefan Breuers, Raja Chatila, Mohamed Chetouani, Daniel Cremers, Vanessa Evers, Michelangelo Fiore, Hayley Hung, Omar A. Islas Ramírez, Michiel Joosse, Harmish Khambhaita, Tomasz Kucner, Bastian Leibe, Achim J. Lilienthal, Timm Linder, Manja Lohse, Martin Magnusson, Billy Okal, Luigi Palmieri, Umer Rafi, Marieke van Rooij, and Lu Zhang. 2016. SPENCER: A Socially Aware Service Robot for Passenger Guidance and Help in Busy Airports. In *Field and Service Robotics*, David S. Wettergreen and Timothy D. Barfoot (Eds.). Springer Tracts in Advanced Robotics, Vol. 113. Springer, Cham, 607–622. doi:10.1007/978-3-319-27702-8_40

[208] John C. Turner, Michael A. Hogg, Penelope J. Oakes, Stephen D. Reicher, and Margaret S. Wetherell. 1987. *Rediscovering the social group: a self-categorization theory*. Basil Blackwell, Oxford, UK.

[209] Laxmi Pandit Vishwakarma, Rajesh Kr Singh, Ruchi Mishra, Denizhan Demirkol, and Tugrul Daim. 2024. The adoption of social robots in service operations: A comprehensive review. *Technology in Society* 76 (March 2024), 102441. doi:10.1016/j.techsoc.2023.102441

[210] Michael L. Walters, Kheng Lee Koay, Dag Sverre Syrdal, Anne Campbell, and Kerstin Dautenhahn. 2013. Companion robots for elderly people: using theatre to investigate potential users' views. In *2013 IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Gyeongju, South Korea, 691–696. doi:10.1109/ROMAN.2013.6628393

[211] Michael L. Walters, Kheng Lee Koay, Dag Sverre. Syrdal, Anne Campbell, and Kerstin Dautenhahn. 2013. Companion Robots for Elderly People: Using Theatre to Investigate Potential Users' Views. In *IEEE RO-MAN 2013*. IEEE, Gyeongju, South Korea, 691–696. doi:10.1109/ROMAN.2013.6628393

[212] Ning Wang, David V. Pynadath, and Susan G. Hill. 2016. Trust calibration within a human-robot team: comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Christchurch, New Zealand, 109–116. doi:10.1109/HRI.2016.7451741

[213] Adam Waytz, Nicholas Epley, and John T. Cacioppo. 2010. Social Cognition Unbound: Insights Into Anthropomorphism and Dehumanization. *Current directions in psychological science* 19, 1 (Feb. 2010), 58–62. doi:10.1177/0963721409359302

[214] Joel Wester, Minha Lee, and Niels Van Berkel. 2023. Moral transparency as a mitigator of moral bias in conversational user interfaces. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) *(CUI '23)*. Association for Computing Machinery, New York, NY, USA, 6 pages. doi:10.1145/3571884.3603752

[215] Tom Williams. 2023. The Eye of the Robot Beholder: Ethical Risks of Representation, Recognition, and Reasoning over Identity Characteristics in Human-Robot Interaction. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23)*. Association for Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3568294.3580031

[216] Tom Williams. 2024. Understanding roboticists' power through matrix guided technology power analysis. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI Companion '24)*. Association for Computing Machinery, New York, NY, USA, 46–56. doi:10.1145/3610978.3640766

[217] Tom Williams, Daniel Ayers, Camille Kaufman, Jon Serrano, and Sayanti Roy. 2021. Deconstructed Trustee Theory: Disentangling Trust in Body and Identity in Multi-Robot Distributed Systems. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21)*. ACM, New York, NY, USA, 262–271. doi:10.1145/3434073.3444644

[218] Tom Williams, Daniel Ayers, Camille Kaufman, Jon Emmanuel Serrano, Shania Jo Runningrabbit, Sayanti Roy, Poulomi Pal, Alexandra Bejarano, and Ryan Blake Jackson. 2020. Identity Performance in Multi-Robot Distributed Systems. In *Proceedings of the Workshop on Human-Robot Interaction for Space Robotics at the 12th International Conference on Social Robotics (ICSR 2020)*. Golden, CO, USA, 1–5. https://rbjackson.github.io/paper_pdfs/williams2020icsr_space.pdf

[219] Cobe Deane Wilson, Danielle Langlois, and Marlena R. Fraune. 2024. Strangers on a Team?: Human Companions, Compared to Strangers or Individuals, are More Likely to Reject a Robot Teammate. *International Journal of Social Robotics* 16, 4 (April 2024), 699–709. doi:10.1007/s12369-024-01133-1

[220] Katie Winkle. 2025. Robots from Nowhere: A Case Study in Speculative Sociotechnical Design and Design Fiction for Human-Robot Interaction. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Melbourne, Australia, 1152–1165. doi:10.1109/HRI61500.2025.10974123

[221] Katie Winkle, Praminda Caleb-Solly, Ute Leonards, Ailie Turton, and Paul Bremner. 2021. Assessing and Addressing Ethical Risk from Anthropomorphism and Deception in Socially Assistive Robots. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21)*. Association for Computing Machinery, New York, NY, USA, 101–109. doi:10.1145/3434073.3444666

[222] Katie Winkle, Ryan Blake Jackson, Alexandra Bejarano, and Tom Williams. 2021. On the Flexibility of Robot Social Identity Performance: Benefits, Ethical Risks and Open Research Questions for HRI. In *Workshop on Robo-Identity: Artificial Identity and Multi-embodiment at the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21)*. HRI 2021, 1–4. https://mirrorlab.mines.edu/publications/winkle2021hri-identity/

[223] Katie Winkle, Donald McMillan, Maria Arnelid, Katherine Harrison, Madeline Balaam, Ericka Johnson, and Iolanda Leite. 2023. Feminist Human-Robot Interaction: Disentangling Power, Principles and Practice for Better, More Ethical HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23)*. Association for Computing Machinery, New York, NY, USA, 72–82. doi:10.1145/3568162.3576973

[224] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. 2021. Boosting Robot Credibility and Challenging Gender Norms in Responding to Abusive Behaviour: A Case for Feminist Robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 29–37. doi:10.1145/3434074.3446910

[225] Katie Winkle and Natasha Mulvihill. 2024. Anticipating the Use of Robots in Domestic Abuse: A Typology of Robot Facilitated Abuse to Support Risk Assessment and Mitigation in Human-Robot Interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 781–790. doi:10.1145/3610977.3634938

[226] Langdon Winner. 1980. Do Artifacts Have Politics? *Daedalus* 109, 1 (1980), 121–136. https://www.jstor.org/stable/20024652

[227] Jie Xu, Kim Le, Annika Deitermann, and Enid Montague. 2014. How different types of users develop trust in technology: A qualitative analysis of the antecedents of active and passive user trust in a shared technology. *Applied Ergonomics* 45, 6 (Nov. 2014), 1495–1503. doi:10.1016/j.apergo.2014.04.012

[228] Jing Yin. 2018. Beyond postmodernism: A non-western perspective on identity. *Journal of Multicultural Discourses* 13, 3 (2018), 193–219. doi:10.1080/17447143.2018.1497640

[229] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16. doi:10.1145/3414685.3417838

[230] Chunpeng Zhai, Santoso Wibowo, and Lily D. Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments* 11, 1 (June 2024), 28. doi:10.1186/s40561-024-00316-7

[231] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. 2025. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3706598.3713429

[232] Yifei Zhu, Colin Brush, and Tom Williams. 2024. Designing augmented reality robot guidance interactions through the metaphors of re-embodiment and telepresence. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Pasadena, CA, USA, 29–36. doi:10.1109/RO-MAN60168.2024.10731244

Received 19 January 2026; accepted 18 March 2026