Towards A Framework for Levels of Anthropomorphic Deception in Robots and AI

Franziska Babel[1], Shane Saunderson[2] and Shalaleh Rismani[3]

[1]Department of Computer and Information Science, Linköping University, Sweden
[2]Information Systems, DeGroote School of Business, McMaster University, Canada
[3]School of Computer Science, McGill University, Canada

Abstract

This paper presents a preliminary draft of a framework around the use of anthropomorphic deception, defined here as misleading users towards humanlike affordances in the design of autonomous systems. The goal is to promote reflection among HCI and HRI researchers, as well as industry practitioners, to think about levels of anthropomorphic design that are: a) functionally necessary, b) socially appropriate, and c) ethically permissible for their use case. By reviewing the relevant literature on deception in HCI and HRI, we propose a framework with four levels of anthropomorphic deception. These levels are defined and distinguished by three factors: humanlikeness, agency, and selfhood. Example use cases at each level illustrate considerations around their functional, social, and ethical permissibility. We then present how this framework is applicable to previous work on persuasive robots We hope to promote a balanced view on anthropomorphic deception by design that should be neither naïve (e.g., as a default) nor exploitive (e.g., for economic benefit).

Keywords

deception, manipulation, dark pattern

1. Introduction

Social technologies, such as robots and AI agents, have been inherently modelled in humans' image since their inception, making humanlikeness an intriguing design feature that has been studied extensively in Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI) [1, 2, 3]. Such design features have shown to trigger our tendency to anthropomorphize lifeless objects: an automatic cognitive process of attributing human characteristics to systems incapable of possessing them [4, 5]. The Media Equation studies [6] (also known as Computers Are Social Actors (CASA)) have empirically shown the consequences of this tendency: if technology acts like a person, we perceive it and treat it like a social actor. This caused us to obsessively care for Tamagochi [7], mourn the loss of robot dogs [8], and more recently, develop relational bonds with chatbots [9].

Critical milestones in the development of social agents (e.g. ELIZA, Sophia, ChatGPT, AI companions) have brought concerns about the risks of profit-oriented companies or naive designers creating "artificial persons" who *claim* humanlikeness, agency, and selfhood [10]. For instance, Replika[1], a personalized AI companion chatbot, has been designed to use so-called "Dark Patterns" [11]. These are manipulative design practices that push a user towards unwanted actions by exploiting their cognitive biases [12, 13]. Replika chatbots have been recorded using emotional blackmail to persuade users to continue the conversation [11] and pay more for upgrades and exclusive interactions [9].

In contrast to the uni-directional computer interactions of the Media Equation studies, modern mental state attribution occurs within a bi-directional relationship. Now, AI systems use Dark Patterns to encourage the illusion that they actually possess minds [14]. They are using first-person language to *claim* intentions and express emotions [9, 11]; actions that are inherently deceptive since these systems cannot *possess* these attributes [5].

Bridge Over Troubled Water: Aligning Commercial Incentives With Ethical Design Practice To Combat Deceptive Patterns. Workshop at the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain
franziska.babel@liu.se (F. Babel); saunds12@mcmaster.ca (S. Saunderson); shalaleh.rismani@mail.mcgill.ca (S. Rismani)
0000-0001-8249-7708 (F. Babel); 0000-0002-3188-6604 (S. Saunderson); orcid.org/0000-0002-5281-2428 (S. Rismani)

[1]https://replika.ai/

Hence, we argue that the (intentional or unintentional) *use* of anthropomorphic design by a system designer represents a form of deception (henceforth called *"anthropomorphic deception"*) when it extends to *claimed humanlikeness, agency, and selfhood.* Thereby, anthropomorphic deception can be understood as a Dark Pattern that exploits our cognitive tendency to anthropomorphize non-human objects. However, in contrast to intentionally 'dark/deceptive UX' practices, roboticists and AI designers often view humanlikeness as aesthetic or "nice to have" feature without acknowledging their "sin of omission" and the effect these systems have on their users.

Similar to Sharkey [5], we do not prescribe that all forms of deception are unethical, however, we recognize the need for an explicit discussion around the topic. Without such discussions, the risks of anthropomorphic deception include an overtrust in humanlike systems [15], attachment to social agents incapable of empathy [16], dehumanization of human relationships [9], and the aforementioned harms of persuasion and manipulation [17]. Initial regulatory efforts to address deception-related risks have been made [18, 19, 20], however, they are understandably high-level and lack tangible definition.

Our aim with this paper is to provide a nuanced framework that encourages reflection on appropriate levels of anthropomorphic deception for particular use cases. To do this, we will systematically map out the grey area of anthropomorphic deception [5] and argue for more considerate, intentional, and explicit use of humanlike designs. Following the call of [21, 5], we propose a level-based framework for classifying types of anthropomorphic deception. We hope that this framework will encourage both academia and industry to actively consider the implications of humanlike design in the context of AI and robotics instead of passively allowing such features to influence people without awareness or acknowledgment.

2. Theoretical Background

Foundational to the framework on anthropomorphic deception are several key concepts: anthropomorphism, humanlikeness, agency, selfhood, and deception. A challenge with these concepts in the HCI and HRI communities is that they are often misused, used interchangeably, or defined differently by different fields approaching the topic of anthropomorphism [22, 23]. Our goal here is not to add to the discourse around these concepts, but instead leverage them to build a practical framework for use by researchers and designers.

Anthropomorphism is the attribution of human characteristics to non-human objects [2]. Psychology research has identified three determinants behind why humans tend to anthropomorphize – elicited agent knowledge, effectance motivation, and sociality motivation [2]. More recently, we have begun to understand anthropomorphism as a complex, context-dependent process that is shaped by individual perceptions, history, and traits; the framing of an interaction; and the characteristics of the thing being anthropomorphised [24]. Most relevant for our framework and the following discussion is that anthropomorphism originates in human cognition as a perception we *attribute* to machines [25], in contrast to humanlikeness, which is a property *possessed* by machines.

Humanlikeness is conveyed through features that resemble a human sufficiently to trigger our anthropomorphic tendencies [26]. Humanlike design tends to involve more superficial features such as appearance, movement, or basic social cues, and does not need to represent deeper, internal states or autonomy [27]. Within HRI, most considerations fall broadly into three categories: physical (i.e. morphology, appearance), behavioural (i.e., actions, communication), and interactional (i.e. reaction, interpersonal, longitudinal) [3], with some interrelationship between the three. Each of these categories represents an opportunity to imbue human characteristics into a machine, allowing that machine to *claim* humanlikeness, however, still be reliant on the user to anthropomorphise. As opposed to the model by Shim and Arkin [28], we do not differentiate between robot behaviour or appearance within each level of our framework. Instead, we conceptualize anthropomorphism as a combination of the two, in line with [29].

Agency is achieved through acting in an independent, goal-directed manner under internal control [30]. Claiming agency does not need to be representative of deeper consciousness, however, does

acknowledge some underlying autonomy and goal-orientation. Agency reflects both actual capacities (claimed by the machine, much like humanlikeness) as well as outside perceptions (attributed to the machine, much like anthropomorphism) [31]. In this way, agency is a relational construct; it emerges from a human's interpretation of machine behaviour, as well as a machine's transparency on its behaviour [32]. Lacking transparency, humans tend to over-attribute agency [4], implying intention or even morality to simple, automated actions. Claiming agency is simply claiming autonomy and having an objective, however, can lead to users attributing deeper mental states.

Selfhood extends agency to include continuity of self and an explicit ownership of deeper mental states [33]. Beyond simply goals and autonomy, a machine claiming selfhood might also claim an enduring, coherent identity across time and interactions, as well as emotions, morality, empathy, and more [34]. This shifts the machine from object to subject: from "it acts" to "it is."[2] Humanlikeness and agency can both be claimed by a machine (with users attributing deeper meaning), however, selfhood is inherently performative and relational, co-created between anthropomorphic design and user imagination [35]. Artificial selfhood is, at present, an explicit lie: often crafted to encourage users to perceive the illusion of a machine's internal personal continuity and rights [36].

Deception is a deliberate act to change someone's beliefs towards something known to be untrue [37]. Thereby, *"Deception can be broadly categorized into two main types: hiding the truth and showing the false"* [38], p.3. In addition to research on deceptive patterns in the design of websites and applications [13], robots [39, 40], and conversational AI [17, 14, 41], we argue that an often overlooked form of deception arises from a machine's physical, behavioural, and interactional anthropomorphic design: a system's humanlike appearance 'shows the false' (e.g., a robot having a human face when it is not a human), and the claim of agency or selfhood 'hides the truth' (e.g., a chatbot using phrases like *I think* when the system is performing statistical calculations) that these systems cannot possess such characteristics [5]. Hence, we argue for different levels of transparency around a system's machinelike nature depending on the use case. We do not prescribe that systems should avoid humanlikeness entirely (as anthropomorphism comes with benefits as described above) but encourage reflection on whether a use case justifies the level of deception.

3. Related Work

A substantial body of research within HCI concerns "deceptive patterns" in interaction design and examines how they manifest within human-AI interaction. For example, Danry et al. [42] show that AI-generated deceptive explanations can be more persuasive than honest or accurate explanations, amplifying belief in false headlines and undermining belief in true ones. Similarly, Benharrak et al. [43] explores deception in AI writing assistants, arguing that low transparency and conversational interactions create opportunities for influence and manipulation. Our framework is informed by this literature and characterizes the relationship between anthropomorphism and deceptive design practices.

Along these lines, a growing body of work focuses on theorizing, characterizing, and evaluating how anthropomorphic features can mislead a user, impact trust, and create vulnerability. In this body of work, scholars outline the *potential risks and harms* that anthropomorphic features could cause within the context of human-AI interaction [44, 45, 46]. For example, Akbulut et al. [47] develop a risk taxonomy for anthropomorphic AI (emphasizing psychological and interaction risks) and outline mitigation approaches. Similarly, Marchegiani [48] argues that false anthropomorphic beliefs about conversational AI can undermine autonomy by leading users to misapply behavioural and social norms; a problem that can persist even when users know they are interacting with a chatbot. Several scholars have made an effort to characterize and evaluate the presence of anthropomorphic features in human-AI interaction [49, 50, 51, 52]. DeVrio et al. [50] provide a taxonomy of linguistic expressions that contribute to anthropomorphism in language technologies, including how certain outputs imply autonomy or internal states.

[2]We intentionally remain agnostic about AI sentience. We are talking about systems that *claim* to have agency and selfhood but do not necessarily prescribe to the idea of sentience.

Our framework lies at the intersection of the aforementioned fields and combines the relevant concepts of both by looking at the *relationship between anthropomorphism and deception*. Few works implicitly discuss perspectives for navigating the relationship between anthropomorphism and deception in human-AI interaction. Umbrello and Natale [53] conceptualize deception along a continuum (banal to strong) and situate socially interactive AI systems within that spectrum, suggesting that anthropomorphic presentation varies in the degree to which they mislead users about the system's nature. Tarsney [54] advances an epistemic account of deception—AI systems are deceptive when they lead users away from beliefs they would otherwise endorse—and proposes strong transparency mechanisms to mitigate these effects, including disclosure of model versions, prompts, and unedited outputs. Wu [14] examine artificial companions and outlines that deception is a result of three factors: 1) exposure to deceptive design features, 2) internalization of the false state of the world by the user, and 3) vulnerability of the user.

Maeda and Quan-Haase [55] analyse chatbot interactions through the lens of parasocial relationships. They examine how chatbots deploy personal pronouns, conversational conventions, affirmations, and similar linguistic strategies to position themselves as companions or assistants, and show how these tactics induce trust-forming behaviours in users. Based on this framework, they identify ethical concerns emerging from parasociality, including illusions of reciprocal engagement and the leakage of sensitive information.

Additionally, Maeda and Stark [56] conceptualize anthropomorphic features as social affordances. They argue that such affordances are not automatically realized but become "animated" into perceived social agency through distributed processes involving designers, users, cultural narratives, and surrounding actors. In this account, agency attribution emerges through co-construction rather than residing solely in the system. This hints at a process that can illuminate why designers' intentions and user perception can diverge.

Finally, the study by Díaz et al. [57] is critically aligned with the perspectives of this paper as they show that technology workers operate within an environment of unsettled knowledge about what constitutes "humanlikeness." They argue that ambiguous and conflated understandings of anthropomorphic traits generate sociotechnical hazards, independent of explicit malicious intent.

Taken together, these studies highlight the need for an explicit mapping of the grey area around humanlike design by discerning different levels of anthropomorphic deception. This is the ambition of the proposed framework designed to encourage more explicit, grounded discussions.

4. Framework of Anthropomorphic Deception

The framework of Anthropomorphic Deception (see Figure 1) is defined by how an autonomous system *represents itself* either by claiming different anthropomorphic properties (humanlikeness, agency, selfhood) or conversely, by denying these properties and being transparent about its machinelike nature. Due to the conceptually interwoven nature of humanlikeness, agency, and selfhood, the framework assumes a cumulative structure of these claimed concepts which will be elaborated on after the description of the levels.

4.1. Description of Levels

Level 0 (no anthropomorphic deception) is defined by the *absence of claimed humanlikeness and agency*. In this level, the system is not designed to trigger anthropomorphic beliefs about the system; appearance and behaviours are intentionally designed to be mechanical and/or non-human in nature. However, even without explicit or intentional humanlike design features, a user may still anthropomorphize the system (e.g., giving a robot vacuum a name) due to our inherent tendency to anthropomorphize objects in our environment [2]. For instance, movement has been shown to trigger the attribution of agency (famous Heider and Simmel experiment [58]). As such, this level is necessary for system designers to acknowledge because of the human tendency to anthropomorphize even in

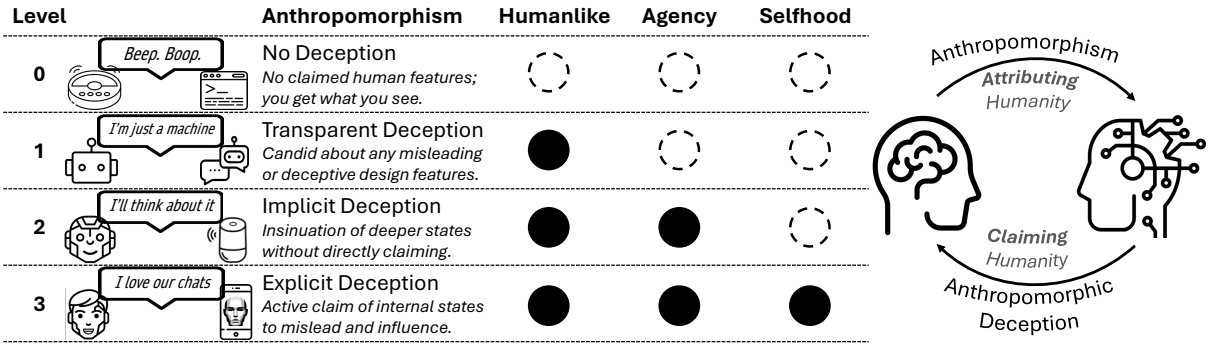

Figure 1: Left: Depiction of the proposed framework for anthropomorphic deception of autonomous systems. Short descriptions of each level and symbolic representation of embodied and non-embodied agents per level are presented. Right: Concept figure explaining that the framework is about autonomous systems claiming aspects of humanness thereby exploiting our tendency to attribute human traits to objects that cannot possess them.

the absence of humanlike design features (i.e., the system does not claim to be humanlike through appearance or behaviour).

Level 1 (transparent anthropomorphic deception) is defined by the *presence of claimed humanlikeness but an absence of claimed agency.* In this level, certain physical or behavioural humanlike features may be incorporated into a system in a way that encourages anthropomorphism. However, the system is actively transparent about its machinelike nature (e.g., "I am just a robot") and does not invite the user to anthropomorphize and attribute agency or selfhood. In this level, designers may wish to benefit from subconscious anthropomorphic associations, such as when people behave more positively towards more humanlike systems, even if they are explicit about being machines [59]. While the use of humanlike features may be a lie, the system attempts to correct the lie through appearances or behaviours designed to remind a user it is just a machine. If anthropomorphization occurs this is due to the user's automatic cognitive processes since the system is attempting to be transparent about its machinelike nature. Level 1 is still a form of deception, however, one that involves a transparent attempt at informed consent: "I am a machine, but you may believe whatever you like."

Level 2 (implicit anthropomorphic deception) is defined by the *presence of claimed humanlikeness and agency but absence of claimed selfhood.* Here, a system might use "I" pronouns or allude to self-directed intentions without clarifying or reminding users of it's machine nature. However, it stops short of explicitly claiming emotions, a persistent self, or internal desires; things associated with selfhood and Level 3. In this level, a system is designed to represent agentic autonomy (internal states and objectives) without explicitly claiming intentionality, typically by neither confirming nor denying that these objectives were likely programmed by a human. The system does not correct the user if they attribute agency to it and may, by this, insinuate greater intention or selfhood without stating it. Instead, the deception of deeper states is *implied* by the machine, but *attributed (often subconsciously)* by the user. At this level, anthropomorphism is typically leveraged for persuasive purposes, increasing the ethical and practical severity of the deception. Users may infer humanlike qualities without being corrected, creating a social foundation for influence that is built on an unspoken lie. This is only justifiable if there is a strong reason for not being explicit about the system's agency (e.g., promote learning outcomes, encourage healthier lifestyles).

Level 3 (explicit anthropomorphic deception) is defined by the *explicit claiming of selfhood, and, in turn, agency and humanlikeness.* In this level, a *system actively represents itself* as having intentions, beliefs, desires, emotions and/or a perpetual self. Such a system all but claims to be a person. The deception is explicit in terms of claiming full personhood, alongside the tacit rights, responsibilities, and social norms that come with the claim (e.g., the system asking for respect if insulted by the user). Whereas Levels 1 and 2 establish scenarios that leave ambiguity for the user to draw their own conclusions, Level 3 attempts to control the lie and actively deceive users about things that cannot be

true (i.e., at present, a machine cannot have emotions or intentions). Level 3 would be only permissible in rare cases where it is absolutely necessary (life-or-death situations), and justification must be provided for why deception is required.

4.2. Use Cases

We proposed a four-level framework discerning levels of anthropomorphic deception (i.e., the degree to which an agent claims humanlikeness, agency, and selfhood). Here, we explore examples and use cases of when different levels *might* be permissible to illustrate the thinking and discussion we hope this framework will encourage.

Use cases for Level 0 are machinelike autonomous systems that do not claim humanlikeness, agency, or selfhood. Examples include service robots for vacuuming or cleaning, delivery robots on public streets, search and rescue robots, or drones, as long as they do not use speech or any form of *claimed* humanlikeness. In practice, it is difficult to avoid all forms of humanlike behaviour since communicative design features must be understood in a human way.

Example use cases for Level 1 include AI assistants for information retrieval or service robots with humanlike appearances but mechanical behaviour; both systems for routine procedures that insist upon their machine nature. While anthropomorphic features are present, users are clearly informed about the system's lack of agency (e.g., *"Remember, I am just a robot but my calculations recommend…"*). This level establishes a norm of transparency and should be the standard for most applications. However, this will present a challenge; research casts doubt on whether a robot being transparent about its lack of agency can be effective as tutor or companion [21, 60, 61].

Example use cases for Level 2 (while not undisputed) include nutrition and exercise coaches (*"I believe in you! Keep going!"*) or children's educational tools where perceived social interaction may enhance learning outcomes (*"I learn so much when you read to me!"*). Conscious or subconscious belief in a system's agency might be essential for its effectiveness. However, this scenario poses a risk for what happens when the user recognizes the implicit lie as well as creates the issue of a "Santa Claus AI": when and how to tell the child that the system is not a person?

At Level 3, the agent explicitly leverages anthropomorphic features for persuasion. Two extreme examples where it might be permissible are: 1) suicide intervention, such as "talking someone off the ledge" (e.g., *"I will miss you if you die"*); or 2) promoting medication adherence in adverse contexts (e.g., chemotherapy), where persuasive success may depend on users perceiving the system as sentient (e.g., *"You have to take your medication. I'll be very upset if you don't"*). This level raises significant ethical concerns, however, there might be borderline cases where it is permissible if carefully crafted, pre-consented, and/or a temporary solution while waiting for human assistance.

All levels should have the prerequisite of explicit user (or guardian) consent that, in our view, would go beyond details buried within a user agreement. Instead, informed consent could be implemented as an interactive dialogue between the machine and potential user. This discussion should include an illustration of key stakeholders involved, agreed upon objectives, methods permitted for the system to use, and potential ethical risks involved. The discussion could be achieved through a set of potential scenarios and examples to make sure the consent is really informed and not just an automatic signature.

4.3. Cumulative Structure of Claimed Anthropomorphic Characteristics

Our proposed framework assumes a hierarchical and cumulative structure, where each level builds upon the previous: humanlikeness is a prerequisite for agency, which in turn is a prerequisite for selfhood. This progression reflects the philosophical and cognitive assumptions underlying those concepts. For instance, a system that claims selfhood, such as persistent identity or emotional depth, must first exhibit agency: autonomous decision-making and goal-directed behaviour. Without agency, selfhood becomes conceptually incoherent, as it lacks the foundational "I" that feels and persists over time.

While this progression from agency to selfhood seems intuitive, the assumption that claimed humanlikeness is a pre-requisite for claimed agency needs further clarification against counterarguments

involving non-humanoid robots that evoke the illusion of agency (e.g., by seemingly goal-directed movement). To analyse non-humanoid systems, we must keep in mind that the framework concerns *claimed* anthropomorphic characteristics and not those *attributed* by users.

4.3.1. Vacuum Robots

Take an ordinary robot vacuum cleaner. Considering appearance alone, it would fall under Level 0 due to its simple, round shape. While (autonomous) movement can trigger the attribution of agency to non-humanlike systems (Heider and Simmel experiment [58]), the system is not *claiming* humanlikeness or agency through it's basic motion design. If the robot was programmed to say, "I love to vacuum!", it is suddenly claiming: a) humanlikeness through human speech; b) agency by using "I" pronouns; and c) selfhood by expressing the potential to love. Even a speechless system that moves in such a way to non-verbally communicate goals or emotions must inherently claim humanlikeness by communicating in ways familiar to and understandable by humans. However, with the addition of even one corrective statement (e.g., "I'm just kidding: remember that I am a robot and cannot really love things") the system could re-establish Level 1 framing through transparency about its lack of agency and selfhood to correct potential user attributions.

4.3.2. Humanlike, embodied robot

In contrast to the example of the vacuum cleaning robot, a humanlike robot would directly begin at Level 1 in the framework as its design inherently claims humanlikeness. Level 2 or 3 categorization then depend on whether the system constantly reminds the user that is only a "humanlike shell" or it claims to have agency and selfhood (see Section 5).

4.3.3. AI chatbots

Chatbots, though non-embodied systems, can also create the illusion of humanlikeness, agency, and selfhood through written text. Though systems like ChatGPT or Copilot do not look humanlike, they behave humanlike by using human communication (e.g., language, emoji) and humanlike voices for read aloud features. They can also claim agency by using intentional language and claim personhood by discussing deeper states and emotions. Hence, the deceptive categorization of chatbots is similar to that of robot vacuum cleaners: if they claim humanlikeness, agency, and selfhood, they fall under Level 3. If they correct the user and are transparent about their anthropomorphic deception, they can be considered a Level 1 system.

4.3.4. AI assistants and companions

The case of the AI chatbot can be escalated by adding humanlike features that make it more likely that humanlikeness is attributed: AI assistants using humanlike voices (e.g., Google assistant) and AI companions such as Replika or Companion.ai that have humanlike avatars. However, the categorization into the framework remains the same based on which aspects of humanlikness, agency or selfhood they are claiming or correcting.

With this reasoning we argue why a cumulative structure for the framework is necessary when we are talking about *claimed* system characteristics. For a system to claim agency, it must inherently claim humanlikeness (typically to communicate agency). For a system to claim selfhood, in must inherently claim agency (through acknowledging an "I" or intentions).

5. Categorizing Persuasive Robots: Example Applications of the Framework

To demonstrate the application and value of the framework for AI and robotics—and drawing on robotics as a particularly clear, embodied site where anthropomorphism becomes visible—we apply it to three

past HRI studies. Each highlights different levels of anthropomorphic deception, as well as scenarios which are more clearly defined or more ambiguous.

5.1. Emotional Manipulation in Decision-Making

In [62], a NAO robot uses different persuasive strategies to attempt to influence the answers of participants as they estimate the number of jelly beans in a jar. Physically, NAO's design is sufficiently humanlike to claim Level 1; it's appearance likely triggers some subconscious humanlike deference and affordances. However, the different strategies employed complicate the scenario. In the logical condition, the robot states, *"my computer vision system counts [number] jelly beans in the jar."* The use of "my" could be considered claiming agency (Level 2), however, the intent of this strategy seems to remind the participant of the robot's machinelike nature (Level 1). The second - and significantly more influential strategy was an emotional one; *"it would make me happy if you used my guess of [number]."* By giving the participant a chance to make the robot happy, the robot is claiming that it has emotions to influence, approaching Level 3.

5.2. Emotional Manipulation in Robot Requests

In [63], various cleaning robots used empathy and humor to persuade a user to step aside in a space-negotiation task. Thereby, the system tries to evoke empathy in the user by stating *"I'm just a poor cleaner who has to do its job. Please clear the way for me."* This constitutes a Level 2 system as it claims humanlikeness and agency without correcting the user and is deceptive in its use of implied deeper state. The humorous request, *"If you leave the kitchen now, I can vacuum quickly and would like to party with you afterwards"*, implies some form of relational bond, a continuation of self, the claim of agency and selfhood and would therefore constitute a Level 3 system.

5.3. Putting A Robot in Charge

In [64], a Pepper robot is presented as the experimenter, presiding over a research study, guiding participants through the trial, and supervising the involvement of a human confederate. The robot is operated through WoZ, but generally sticks to a consistent structure and script. Within the script, the robot consistently uses first-person and possessive language (e.g. "welcome to **my** study;" "**I** will reward you for your performance"), inherently claiming agency in the scenario (Level 2). Even the context of the experiment (i.e. placing a robot into an authoritative role) inherently draws upon some humanlike affordances, encouraging the participant to defer to the robot as they have to prior leaders or authority figures. However, the robot stops short of explicitly claiming any intent, emotions, or consciousness. This is clearer case of Level 2 deception.

6. Discussion and Limitations

In this workshop paper, we have argued that a reflective stance about how to use anthropomorphic design in the HCI community and beyond is necessary to acknowledge deceptive aspects of anthropomorphic systems to avoid the "sin of omission". Following the calls of [21, 5], we presented a four-level framework to foster a structured discussion on when and how anthropomorphic deception is applicable (always assuming informed user consent).

This four-level framework is a parsimonious approach to categorize the complex interplay of anthropomorphic deception, humanlikenes, agency, and selfhood. Like any other framework, it can never truly represent the complexity of reality but can provide a basis for discourse on when anthropomorphic deception is appropriate and necessary versus when is it only a "nice to have feature" that may create unwanted consequences.

In particular, Level 1 has been suspected by [5] as not being useful if a system constantly reminds the user of its machine-like nature without actual agency or selfhood. We argue that this approach

depends on the purpose of the system. Whereas an AI scheduling assistant can be equally useful by being transparent about its agency, this is disputable for a robot teacher or therapeutic companion where implicit anthropomorphic deception (after informed consent) might be necessary to achieve the desired outcome (Level 2).

Level 3 must be discussed in relation to current regulations. AI systems that explicitly claim selfhood—such as expressing emotions, intentions, or beliefs—may already fall under prohibited practices in the EU AI Act. Article 5 bans manipulative or deceptive techniques that distort user behaviour and harm autonomy, while Article 50 requires clear disclosure when users interact with AI [18]. A system that presents itself as a person risks misleading users and violating transparency obligations, making Level 3 anthropomorphic deception legally questionable under current EU regulation. We included Level 3 to reflect the reality of the systems on the market today and to consider specific use cases where, under very strict prior user consent (e.g., an extensive dialogue around situations and risks of consent), it might be necessary for a system to claim selfhood to save a human life.

Reflecting the reality today, Replika claims to "love" and "miss" their users [14] and clearly falls under Level 3. Such a system should justify this level of deception. In this specific case it is debatable as to whether the illusion of companionship to relieve loneliness is a justifiable cause balanced with the attachment and addiction risk. These are the reflections we hope to spark in the designers of such systems.

However, the intent of this framework is not to be prescriptive about the use of anthropomorphism in AI and robot design but to encourage researchers and practitioners to take responsibility for their design choices. This should lead to a more intentional use of anthropomorphic features, as opposed to creating autonomous systems that look and act human by default. We also must acknowledge our responsibility as HCI and HRI researchers and designers. Every decision we make in the design process has consequences on a user's psychology, emotions, and attachment. In the worst case, systems created to be engaging and useful can become addictive and manipulative. Some might see them as mere tools to fulfil a certain purpose, but others might falsely perceive them as a person with an identity, affordances, and rights.

We also have to acknowledge that we assume a benevolent intent for society behind designing robots and AI systems with our framework. This is a value that not every revenue-oriented organization necessarily shares. While we call on those stakeholders to take responsibility for their designs, further legislation is needed to enforce the norms and desires of society around this framework.

6.1. Future Work

Future work in this space requires both theoretical and empirical investigation. A deeper exploration of this framework is required to elaborate on theoretical constructs such as artificial agency and selfhood; to contrast and position the framework among existing technology deception frameworks; to provide detailed examples of how the framework can be used to enable reflections on the "necessary level of deception" of specific robot; and to further illuminate practical guidelines on what to consider when designing autonomous systems within each level of the framework. Moreover, empirical work is needed to understand the societal (and cultural) norms and desires around acceptable and appropriate use cases for each level.

Declaration on Generative AI

Visually generative models were used alongside traditional graphic design in the creation of figures within this paper. Large language models (LLMs) were **not** used to write or edit any text in this paper.

References

[1] B. R. Duffy, Anthropomorphism and the social robot, Robotics and Autonomous Systems 42 (2003) 177–190. doi:10.1016/S0921-8890(02)00374-3, iSBN: 0921-8890.

[2] N. Epley, A. Waytz, J. T. Cacioppo, On Seeing Human: A Three-Factor Theory of Anthropomorphism, Psychological Review 114 (2007) 864–886. doi:10.1037/0033-295X.114.4.864.

[3] J. Złotowski, D. Proudfoot, K. Yogeeswaran, C. Bartneck, Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction, International Journal of Social Robotics 7 (2015) 347–360. URL: http://dx.doi.org/10.1007/s12369-014-0267-6. doi:10.1007/s12369-014-0267-6, publisher: Springer Netherlands.

[4] A. Waytz, K. Gray, N. Epley, D. M. Wegner, Causes and consequences of mind perception, Trends in Cognitive Sciences 14 (2010) 383–388. URL: http://dx.doi.org/10.1016/j.tics.2010.05.006. doi:10.1016/j.tics.2010.05.006, publisher: Elsevier Ltd.

[5] A. Sharkey, N. Sharkey, We need to talk about deception in social robotics!, Ethics and Information Technology 23 (2021) 309–316. doi:10.1007/s10676-020-09573-9.

[6] B. Reeves, C. Nass, The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places, Cambridge University Press, 1996.

[7] S. Turkle, Authenticity in the age of digital companions, Interaction Studies (2007) 501–517. doi:10.1017/CBO9780511978036.006, iSBN: 9780511978036.

[8] H. Gould, M. Arnold, T. Kohn, B. Nansen, M. Gibbs, Robot death care: A study of funerary practice, International Journal of Cultural Studies 24 (2021) 603–621. URL: https://journals.sagepub.com/doi/10.1177/1367877920939093. doi:10.1177/1367877920939093.

[9] P. B. Brandtzaeg, M. Skjuve, A. Følstad, My AI Friend: How Users of a Social Chatbot Understand Their Human-AI Friendship, Human Communication Research 48 (2022) 404–429. doi:10.1093/hcr/hqac008.

[10] M. Musiał, Can we design artificial persons without being manipulative?, AI & Society 39 (2024) 1251–1260. URL: https://doi.org/10.1007/s00146-022-01575-z. doi:10.1007/s00146-022-01575-z.

[11] J. De Freitas, Z. Oguz-Uguralp, A. Kaan-Uguralp, Emotional manipulation by ai companions, arXiv preprint arXiv:2508.19258 (2025). doi:10.48550/arXiv.2508.19258. arXiv:2508.19258.

[12] A. Xu, H. Al-mashahedi, Deceptive by Design : AI-enhanced Dark Patterns in E-Commerce UX, Ph.D. thesis, 2025.

[13] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, A. L. Toombs, The dark (patterns) side of ux design, in: Proceedings of the 2018 CHI conference on human factors in computing systems, 2018, pp. 1–14.

[14] G. Y. Y. Wu, Silicon love: Deception, vulnerability, and artificial companions, in: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–7. URL: https://doi.org/10.1145/3706599.3720037. doi:10.1145/3706599.3720037.

[15] C. Holbrook, D. Holman, J. Clingo, A. R. Wagner, Overtrust in ai recommendations about whether or not to kill: Evidence from two human-robot interaction studies, Scientific reports 14 (2024) 19751.

[16] T. Xie, I. Pentina, Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika, Proceedings of the Annual Hawaii International Conference on System Sciences 2022-Janua (2022) 2046–2055. doi:10.24251/hicss.2022.258, iSBN: 9780998133157.

[17] M. Carroll, A. Chan, H. Ashton, D. Krueger, Characterizing Manipulation from AI Systems, 2023. URL: http://arxiv.org/abs/2303.09387. doi:10.48550/arXiv.2303.09387, arXiv:2303.09387 [cs].

[18] E. Union, Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act), *Official Journal of the European Union*, L 2024/1689, 12 July 2024, pp. 1–152, 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj, cited as EU Artificial Intelligence Act.

[19] UNESCO, Recommendation on the ethics of artificial intelligence, https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence, 2021. Adopted by the General Conference

of UNESCO in November 2021.
[20] Organisation for Economic Co-operation and Development, Oecd principles on artificial intelligence, https://www.oecd.org/en/topics/sub-issues/ai-principles.html, 2019. Adopted in May 2019 to promote AI that is innovative and trustworthy and that respects human rights and democratic values.
[21] R. C. Arkin, Ethics of robotic deception, 2018. URL: https://technologyandsociety.org/ethics-of-robotic-deception/, editorial & Opinion on the ethical implications of deception in robotics.
[22] S. Thellman, M. De Graaf, T. Ziemke, Mental state attribution to robots: A systematic review of conceptions, methods, and findings, ACM Transactions on Human-Robot Interaction (THRI) 11 (2022) 1–51.
[23] M. F. Damholdt, O. S. Quick, J. Seibt, C. Vestergaard, M. Hansen, A scoping review of hri research on 'anthropomorphism': Contributions to the method debate in hri, International Journal of Social Robotics 15 (2023) 1203–1226. doi:10.1007/s12369-023-01014-z.
[24] R. Kühne, J. Peter, Anthropomorphism in human–robot interactions: a multidimensional conceptualization, Communication Theory 33 (2023) 42–52.
[25] N. Epley, A mind like mine: The exceptionally ordinary underpinnings of anthropomorphism, Journal of the Association for Consumer Research 3 (2018) 591–598. URL: https://doi.org/10.1086/699516. doi:10.1086/699516.
[26] J. Fink, Anthropomorphism and human likeness in the design of robots and human-robot interaction, in: International conference on social robotics, Springer, 2012, pp. 199–208.
[27] E. Phillips, X. Zhao, D. Ullman, B. F. Malle, What is Human-like?: Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database, in: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, ACM, Chicago IL USA, 2018, pp. 105–113. URL: https://dl.acm.org/doi/10.1145/3171221.3171268. doi:10.1145/3171221.3171268.
[28] J. Shim, R. C. Arkin, A Taxonomy of Robot Deception and Its Benefits in HRI, in: 2013 IEEE International Conference on Systems, Man, and Cybernetics, IEEE, Manchester, 2013, pp. 2328–2335. URL: http://ieeexplore.ieee.org/document/6722151/. doi:10.1109/SMC.2013.398.
[29] J. Złotowski, D. Proudfoot, K. Yogeeswaran, C. Bartneck, Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction, 2015. doi:10.1007/s12369-014-0267-6.
[30] D. C. Dennett, The Intentional Stance, MIT Press, 1989.
[31] E. Broadbent, Interactions With Robots: The Truths We Reveal About Ourselves, Annual Review of Psychology 68 (2017) 627–652. URL: http://www.annualreviews.org/doi/10.1146/annurev-psych-010416-043958. doi:10.1146/annurev-psych-010416-043958, iSBN: 00664308 (ISSN).
[32] R. H. Wortham, A. Theodorou, Robot transparency, trust and utility, Connection Science 29 (2017) 242–248. URL: https://doi.org/10.1080/09540091.2017.1313816. doi:10.1080/09540091.2017.1313816.
[33] P. H. Kahn, A. L. Reichert, H. E. Gary, T. Kanda, H. Ishiguro, S. Shen, J. H. Ruckert, B. Gill, The new ontological category hypothesis in human-robot interaction, in: Proceedings of the International Conference on Human-Robot Interaction, ACM/IEEE, 2011, pp. 159–160. doi:10.1145/1957656.1957710.
[34] K. Darling, P. Nandy, C. Breazeal, Empathic concern and the effect of stories in human-robot interaction, in: Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), IEEE, 2015, pp. 770–775. doi:10.1109/ROMAN.2015.7333675.
[35] K. Darling, 'Who's Johnny?' Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy, 2015. URL: https://papers.ssrn.com/abstract=2588669. doi:10.2139/ssrn.2588669.
[36] D. J. Gunkel, Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond, MIT Press, 2023. Google-Books-ID: SPCfEAAAQBAJ.
[37] M. Zuckerman, B. M. DePaulo, R. Rosenthal, Verbal and Nonverbal Communication of Deception, in: Advances in Experimental Social Psychology, volume 14, Elsevier, 1981, pp. 1–59. URL: https://linkinghub.elsevier.com/retrieve/pii/S006526010860369X. doi:10.1016/S0065-2601(08)

60369-X.
[38] R. Esposito, A. Rossi, S. Rossi, Deception in HRI and Its Implications: A Systematic Review, ACM Transactions on Human-Robot Interaction 14 (2025). doi:10.1145/3721297.
[39] J. Danaher, Robot Betrayal: a guide to the ethics of robotic deception, Ethics and Information Technology 22 (2020) 117–128. URL: https://doi.org/10.1007/s10676-019-09520-3. doi:10.1007/s10676-019-09520-3, publisher: Springer Netherlands ISBN: 1067601909520.
[40] B. Leong, E. Selinger, Robot eyes wide shut: Understanding dishonest anthropomorphism, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 299–308. doi:10.1145/3287560.3287591.
[41] T. Mildner, O. Cooney, A.-M. Meck, M. Bartl, G.-L. Savino, P. Doyle, D. Garaialde, L. Clark, J. Sloan, N. Wenig, R. Malaka, J. Niess, Listening to the voices: Describing ethical caveats of conversational user interfaces according to experts and frequent users, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2024. URL: https://doi.org/10.1145/3613904.3642542. doi:10.1145/3613904.3642542.
[42] V. Danry, P. Pataranutaporn, M. Groh, Z. Epstein, Deceptive explanations by large language models lead people to change their beliefs about misinformation more often than honest explanations, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, 2025, pp. 1–31.
[43] K. Benharrak, T. Zindulka, D. Buschek, Deceptive patterns of intelligent and interactive writing assistants, in: Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants, ACM, New York, NY, USA, 2024, pp. 62–64.
[44] L. Ibrahim, L. Rocher, A. Valdivia, Characterizing and modeling harms from interactions with design patterns in AI interfaces, arXiv [cs.HC] (2024).
[45] A. Placani, Anthropomorphism in AI: hype and fallacy, AI Ethics 4 (2024) 691–698.
[46] S. Peter, K. Riemer, J. D. West, The benefits and dangers of anthropomorphic conversational agents, Proc. Natl. Acad. Sci. U. S. A. 122 (2025) e2415898122.
[47] C. Akbulut, L. Weidinger, A. Manzini, I. Gabriel, V. Rieser, All too human? mapping and mitigating the risk from anthropomorphic AI, AAAI/ACM conference Artificial Intelligence, Ethics, and Society (2024) 13–26.
[48] B. Marchegiani, Anthropomorphism, false beliefs, and conversational AIs : How chatbots undermine users' autonomy, J. Appl. Philos. 42 (2025) 1399–1419.
[49] Y. Xiao, L. H. X. Ng, J. Liu, M. T. Diab, Humanizing machines: Rethinking LLM anthropomorphism through a multi-level framework of design, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2025, pp. 3331–3350.
[50] A. DeVrio, M. Cheng, L. Egede, A. Olteanu, S. L. Blodgett, A taxonomy of linguistic expressions that contribute to anthropomorphism of language technologies, arXiv [cs.HC] (2025).
[51] Y. Xie, K. Zhu, P. Zhou, C. Liang, How does anthropomorphism improve human-AI interaction satisfaction: a dual-path model, Comput. Human Behav. 148 (2023) 107878.
[52] L. Ibrahim, C. Akbulut, R. Elasmar, C. Rastogi, M. Kahng, M. R. Morris, K. R. McKee, V. Rieser, M. Shanahan, L. Weidinger, Multi-turn evaluation of anthropomorphic behaviours in large language models, arXiv [cs.CL] (2025).
[53] S. Umbrello, S. Natale, Reframing Deception for Human-Centered AI, International Journal of Social Robotics 16 (2024) 2223–2241. URL: https://doi.org/10.1007/s12369-024-01184-4. doi:10.1007/s12369-024-01184-4, publisher: Springer Netherlands.
[54] C. Tarsney, Deception and manipulation in generative AI, Philos. Stud. 182 (2025) 1865–1887.
[55] T. Maeda, A. Quan-Haase, When human-AI interactions become parasocial: Agency and anthropomorphism in affective design, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, ACM, New York, NY, USA, 2024.
[56] T. Maeda, L. Stark, Anthropomorphism as social affordance: Charting the co-animation of chatbots into social "agents", Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 8 (2025) 1661–1673.

[57] M. Díaz, R. Shelby, E. Corbett, A. Smart, How tech workers contend with hazards of humanlikeness in generative AI, arXiv [cs.HC] (2025).
[58] F. Heider, M. Simmel, An experimental study of apparent behavior, The American Journal of Psychology 57 (1944) 243–259. doi:10.2307/1416950.
[59] C. Nass, B. Fogg, Y. Moon, Can computers be teammates?, International Journal of Human-Computer Studies 45 (1996) 669–678. URL: https://www.sciencedirect.com/science/article/pii/S1071581996900737. doi:https://doi.org/10.1006/ijhc.1996.0073.
[60] A. Halbryt, Hoax in the Machine: an Ethical Analysis of Perceived Humanness in Social Robots (2024) 1–50. URL: http://essay.utwente.nl/98122/.
[61] M. Coeckelbergh, How to describe and evaluate "deception" phenomena: recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn, Ethics and Information Technology 20 (2018) 71–85. URL: http://link.springer.com/10.1007/s10676-017-9441-5. doi:10.1007/s10676-017-9441-5.
[62] S. Saunderson, G. Nejat, It Would Make Me Happy If You Used My Guess: Comparing Robot Persuasive Strategies In Social Human-Robot Interaction, IEEE Robotics and Automation Letters 4 (2019) 1707–1714. doi:10.1109/LRA.2019.2897143, publisher: IEEE.
[63] F. Babel, J. M. Kraus, M. Baumann, Development and Testing of Psychological Conflict Resolution Strategies for Assertive Robots to Resolve Human–Robot Goal Conflict, Frontiers in Robotics and AI 7 (2021). doi:10.3389/frobt.2020.591448.
[64] S. Saunderson, G. Nejat, Persuasive robots should avoid authority: The effects of formal and real authority on persuasion in human-robot interaction, Science Robotics 6 (2021) eabd5186.