Great Chain of Agents: The Role of Metaphorical Representation of Agents in Conversational Crowdsourcing

JI-YOUN JUNG, Knowledge and Intelligence Design, Delft University of Technology, The Netherlands
SIHANG QIU, Web Information Systems, Delft University of Technology, The Netherlands
ALESSANDRO BOZZON, Knowledge and Intelligence Design, Delft University of Technology, The Netherlands
UJWAL GADIRAJU, Web Information Systems, Delft University of Technology, The Netherlands

Conversational agents are being widely adopted across several domains to serve a variety of purposes ranging from providing intelligent assistance to companionship. Recent literature has shown that users develop intuitive folk theories and a metaphorical understanding of conversational agents (CAs) due to the lack of a mental model of the agents. However, investigation of metaphorical agent representation in the HCI community has mainly focused on the human level, despite non-human metaphors for agents being prevalent in the real world. We adopted Lakoff and Turner's *'Great Chain of Being'* framework to systematically investigate the impact of using non-human metaphors to represent conversational agents on worker engagement in crowdsourcing marketplaces. We designed a text-based conversational agent that assists crowd workers in task execution. Through a between-subjects experimental study (N = 341), we explored how different human and non-human metaphors affect worker engagement, the perceived cognitive load of workers, intrinsic motivation, and their trust in the agents. Our findings bridge the gap of how users experience CAs with non-human metaphors in the context of conversational crowdsourcing.

Additional Key Words and Phrases: Conceptual metaphors; Conversational agent; Crowdsourcing; Engagement; Trust; Great chain of being; Human-agent interaction; Human-AI interaction

ACM Reference Format:
Ji-Youn Jung, Sihang Qiu, Alessandro Bozzon, and Ujwal Gadiraju. 2022. Great Chain of Agents: The Role of Metaphorical Representation of Agents in Conversational Crowdsourcing. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 32 pages. https://doi.org/10.1145/3491102.3517653

1 INTRODUCTION

Conversational interfaces have been argued to have advantages over conventional GUIs due to facilitating a more human-like interaction [74]. The rise in popularity of conversational AI agents has enabled humans to interact with machines more naturally [39]. In addition, people have a growing familiarity with conversational interactions mediated by technology due to the widespread use of mobile devices and messaging services. This has contributed to a steep rise in the use of conversational agents across several domains [41, 46, 51, 101]. Recent work has also shown that crowd workers in microtask marketplaces can complete various human intelligence tasks (HITs) using conversational interfaces, resulting in a similar output quality compared to traditional Web interfaces while exhibiting more engagement and satisfaction [69, 85, 86, 88].

Research in the HCI community has paid attention to the metaphorical representation of artificial intelligence (AI) agents to improve human-agent interaction and inform future design choices. For instance, Khadpe *et al.* revealed that

Manuscript submitted to ACM

metaphors with different degrees of perceived warmth and competency shape pre-use user expectation towards the agent, which leads to disparate effects on intention to adopt, desire to co-operate, and intention to try out a system [56]. Most prior works have only investigated agent metaphors at a human level. However, non-human representation of conversational agents is widespread in the real world, often in the form of a robot (e.g., Woebot[1]) or a bird (e.g., on Duolingo[2] or Stanford's QuizBot[3]). Social simulation games, such as Animal Crossing[4] and its vast commercial success, show that conversation with animal-looking agents who "act like a human" can be as engaging as interacting with agents resembling humans. This can be explained on the grounds that such agents representing animals still encompass dimensions that evoke anthropomorphic perceptions, such as facial expression, body language, or use of human language followed by verbal social cues (joking, response time, body gestures, etc.).

Although conversation, in general, is a highly human trait, metaphorical understanding is often conducted semantically and is often used with non-human attributes. For example, metaphors like '*The man is a wolf*' or '*Achilles is a lion*' have a **<A is a B>** format. Here **A** is explained with certain instinctive traits of a **B**. Likewise, people might have different perceptions of Animal Crossing non-player characters (NPC), between an NPC represented as an eagle or a lion and an NPC represented as a raccoon or a monkey. However, there is a limited understanding of how non-human metaphorical representations of conversational agents shape user engagement, perceived cognitive load, intrinsic motivation, and trust in the agent.

To address the knowledge gap, in this paper, we adopted the '*Great Chain of Being* (GCoB)' framework by Lakoff and Turner [52, 64] (Figure 2 (b)). By doing so, we introduce a new lens to frame and design non-human agent metaphors. This can help agent designers base their design choices in a conceptually more structured way that extends the current understanding of agent metaphors.

GCoB metaphor is composed of hierarchical scales of god, human, animal, plant, and inorganic object. According to Lakoff and Turner, the GCoB metaphor is one of the unconscious cognitive models that we as humans use to understand and categorize the world around us [64]. GCoB's hierarchical and vertical scale suggests that entities in any level have all the properties that any lower levels possess, in addition to their distinctive property that a lower level does not have. Therefore, it is helpful to understand the complex faculties of human beings in terms of the lower-level property. Furthermore, based on the previous findings from Khapede *et al.* that metaphors with higher perceived competence resulted in lower intention to adopt a given system [56], hypothesizing that such a vertical scale also links to the level of competence and warmth, it is interesting to see if agents designed with metaphors from the higher level have a different impact on users.

In addition, humans subconsciously maintain separate schemas that characterize our knowledge about people from schemas of the physical world (p.162, [64]). The GCoB metaphor allows us to link such disparate schemas. For example, if someone were to be called a rock, most people would interpret the person as strong, persistent, or stubborn. Such inferences can be drawn from people's understanding of the characteristics of a rock, for example, that rock is usually firm, and if a rock is large, the rock is likely heavy and hard to move around. Such examples can also be found in the famous '*Computer is a Desktop*' metaphor (Figure 2 (a)), where it helped people comprehend the unfamiliar concept (Computer) as attributes of the well-understood concepts (desktop). This GCoB framework was developed to extend Lakoff's well-known work on the Conceptual Metaphor theory [63]. Conceptual metaphor theory treats

[1]https://woebothealth.com/
[2]https://www.duolingo.com/log-in
[3]https://hci.stanford.edu/research/smartprimer/projects/quizbot.html
[4]https://www.nintendo.com/games/detail/animal-crossing-new-horizons-switch/

metaphors as conceptual rather than purely linguistic entities, which involves a systemic projection between two mental representations (conceptual domains).

Research on embodied conversational agents (ECAs) has explored the effect of anthropomorphism on users. ECAs have been proposed to handle multimodal input and output, the production and interpretation of gestures and emotions, and the development of avatars and talking heads [73]. This can be shown by commercial software such as Apple's Memoji, where people can create personalized "animoji (animated emoji)" that has a shape of a dog, monkey, bear, and so on. Such software is developed in order to support expressive digital communication. Drawing inspiration, one can design conversational agents and utilize non-human metaphors to explain the agent and adjust users' expectations towards the agent.

We adopted the lens of conversational microtask crowdsourcing in this work, where improving worker satisfaction remains a challenge [57]. With a growing demand around human input due to the rapid advancement of automation, robotics, and AI, designing human intelligence tasks (HITs) that are engaging is a crucial research topic. However, HITs on microtasking platforms can be painfully monotonous, leading to high drop-out, rejection, and task abandonment rates [25, 44, 45, 72]. In order to improve this, recent literature suggested using a conversational interface to conduct HITs of different types (e.g., image transcription, information finding, sentiment analysis, image classification) in microtasking platforms. Results showed that a conversational interface improved perceived worker engagement and significantly high worker retention while maintaining output quality compared to conventional web task execution interfaces [86, 87].

Furthermore, in a domain where users (i.e., crowd workers) are primarily motivated by monetary rewards, recent work has shown the potential of using worker avatars to improve worker engagement and experience [84]. In this context of conversational crowdsourcing, it is exciting to study and better understand the role and potential of metaphorical representations of a conversational agent. Thus, we investigate the following research question:

> **RQ:** How do different metaphorical representations of a conversational agent impact worker engagement, perceived cognitive load, intrinsic motivation, and trust in conversational microtask crowdsourcing?

To address the research question, we developed a text-based conversational interface using TickTalkTurk [88]. We carried out a between-subjects study spanning 12 experimental conditions (6 metaphors × 2 task types), and recruited 341 workers from Prolific[5] – a crowdsourcing marketplace. We present empirical findings of different agent metaphors derived from five hierarchical categories based on the Great Chain of Being, and their impact on worker engagement, perceived cognitive task load, intrinsic motivation, and trust. We found that metaphorical representations derived from the Great Chain of Being's hierarchical categories can affect worker engagement, intrinsic motivation, and cognitive task load. We show that there is a trade-off in terms of using different metaphors. For instance, using an inorganic object metaphor (book) can significantly reduce the cognitive workload but negatively affect intrinsic motivation. Our study highlights the importance of choosing an appropriate metaphor to represent a conversational agent when designing crowdsourcing tasks.

Original Contributions. This paper makes the following contributions:

- We find evidence for the trade-offs between using different representations from the Great Chain of Being's levels, and provide design implications for conversational crowdsourcing and conversational agents.

[5]https://www.prolific.co

- We propose a method to systematically analyze non-human metaphorical representations in a conceptually structured framework that can be used by agent designers and by researchers studying human-agent interaction.
- We enrich the current discourse on users' metaphorical understanding of conversational agents and address the knowledge gap on metaphorical agent representations.

2 BACKGROUND AND RELATED WORK

2.1 Conversational Agents Explained with Conceptual Metaphors

Over the last few years, conversational agents (CA) have been increasingly studied in the HCI community due to the human-like interaction that they facilitate, touted as "the next natural form of HCI" [67]. In the wake of such CAs, conversational user interfaces are being widely embedded across several domains in personal technologies and devices.

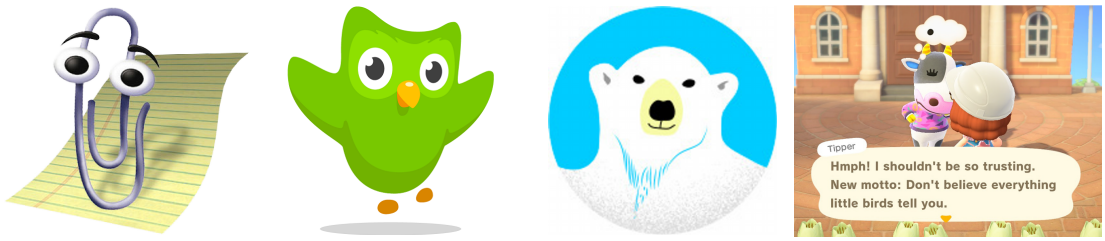
The term conversational agent has various connotations ranging from voice interfaces, virtual companions, virtual agents, autonomous agents, embodied conversational agents (ECAs), to chatbots. In this research, we focus on chatbots [39, 104]. Chatbots have been shown to predict users' attitudes by checking how users greet bots [65], enhance the collaborative experience of users [6], train non-expert users for skill acquisition [4], or improve patient engagement for health literacy [9].

Despite the promising possibilities of CAs in improving user experiences [74], people still lack mental models of such agents and fail to bridge the gap between user expectations and agent operation [67], which also can be explained by Norman's *'gulf of execution'* [77]. Gulf of execution refers to a gap between what a user intends to do and what the system requires the users to do. Norman argues that the gulf of execution should be as small as possible to achieve higher usability. In an attempt to understand user behavior towards AI systems and CAs, recent literature has argued that people develop *folk theories* [36, 82] to reason about cyber-social systems [24, 28, 96]. The folk theory argues that users form an intuitive and informal understanding of the system, for example, how *a helicopter*, *gravity*, or *artificial intelligence* works, to explain their outcomes and consequences. Therefore, users' understanding is often imprecise. Their mental representation of technology is an implicit collection of beliefs rather than a blueprint of inputs and outputs [28]. In a similar vein, the conceptual metaphor theory provides a framework that can potentially reveal how users interact with CAs.

Metaphor has been one of the central themes of the design research discipline [10]. Previous studies have suggested metaphors can not only be a tool for designers to help users understand the system better (e.g., '*Computer is a Desktop*') [68, 92], but also help designers to frame the problem using the Generative Metaphor framework [91]. Lakoff and Johnson [63] argue in their work "Metaphors We Live By [63]" that our conceptual system is fundamentally metaphoric. It is essentially human to understand and explain the world, concept, or ideas by "cross-domain mapping" the target to the source (Figure 2 (a)). Adopting this Conceptual Metaphor framework, a recent study by Khadpe *et al.* investigated how different human metaphors, based on their perceived warmth and competence, influence their expectations, intention to adopt an AI system, and desire to co-operate with the system [56]. Similar to Khadpe *et al.*'s findings, although not in the context of metaphorical representation but also using perceived warmth and competence, Gilad *et al.* identified a primacy for warmth over competence when interacting with AI systems [37]. Our work complements these findings and introduces a new lens to investigate the metaphorical understanding of users via a cross-domain mapping on different hierarchical metaphors in the context of conversational microtask crowdsourcing.

2.2 Non-human Metaphors and Conversational Agents

To simulate human-human dialogue, prior works related to embodied conversational agents (ECAs) have explored the degree of anthropomorphism and its effect on user trust [17, 20, 102], user satisfaction [55, 83], sympathetic social behavior [34, 99], telepresence [79], and user interaction [14]. A recent study by Kuzminykh *et al.* demonstrated that users consistently perceive such agents based on the agent's anthropomorphized behavioral and visual perceptions [62]. As such, the works mentioned here and most recent research efforts in the HCI community have primarily focused on human representations. However, people's metaphorical understanding of an agent is not solely manifested in a human form [37, 42, 64]. Real-world use cases provide an abundance of clear examples of how users can metaphorically understand an agent when the agent is in non-human form (Figure 1). These examples "provide a provocative contrast with the feminine yet disembodied virtual assistants of today" (e.g., Amazon's Alexa, Apple's Siri) [7].



(a) Microsoft Office ***Clippy*** **(b)** Duolingo ***Duo*** **(c)** UX Collective ***UX Bear*** **(d)** Animal Crossing ***Villagers***

Fig. 1. Popular examples of non-human agents employed in different contexts. (a) Clippy used in Microsoft's Office 97 until Office 2017, (b) Duo, pedagogical agent from Duolingo, (c) UX bear, a chatbot from the UX collective (https://uxchat.me/), (d) A Villager in Nintendo's social simulation game Animal Crossing.

Lakoff and Turner (2009) adopted the *Great Chain of Being* metaphor and argued that "when the hierarchy of the basic Great Chain is combined [...] we get a more elaborated, hierarchical folk theory of forms of being and how they behave" (p.171 [64]). The authors discussed how the Great Chain of Being metaphor is "a tool of great power and scope" (p.172 [64]) because it explains a tendency of people to comprehend complex concepts, subjects, or objects such as

(a) Anatomy of a metaphor with an example **(b)** Great Chain of Being framework

Fig. 2. (a) Anatomy of a metaphor with the 'computer is a desktop' metpahor, and (b) the Great Chain of Being represented schematically.

general human traits in terms of well-understood non-human attributes, and vice versa. The extended version of the framework [61] can be represented as illustrated in Figure 2 (b).

GCoB's hierarchical structure implies that entities corresponding to a higher level of the GCoB have all the qualities that the lower levels possess, in addition to their distinct qualities. For example, in the "*A man is a wolf*" metaphor, a human (man) is associated with possessing an animal-like aggressiveness. The GCoB metaphor allows us to project characteristics of well-known or understood from one category onto another. For example, humans can call a glass of red wine "suggestive" or "romantic," exploiting a higher-order feature to explain its taste or surrounding context. On the other hand, the metaphor "*A lie has no legs*" projects an abstract concept onto a physical trait, implying that a lie has nothing to support it or 'no legs' to stand on.

If an entity corresponds to a higher level of the GCoB metaphor, it is unclear whether one would attribute a higher competence to the entity. This is a difficult question to answer since an entity's competence will differ based on the entity's task. For example, a wolf will perform better in the task of hunting than an inexperienced graduate student, but the graduate student will be better at writing an academic paper. However, when non-human metaphors are used in a conversational user interface, they speak fluent human language, which anthropomorphizes the non-human agent. Therefore, we can hypothesize that the perceived competency of an agent might correspond to the hierarchy of the GCoB. To this end, the GCoB metaphor can be used as a framework to analyze anthropomorphized conversational agents for their warmth and competence.

Nowak and Biocca reported that when a virtual agent's image was more unusual and iconic (less anthropomorphic), people got more immersed in the virtual environment and found less anthropomorphic images to be more credible and likable than the more anthropomorphic images [78, 79]. This result poses a question around whether non-human representations of agents –by virtue of being more "iconic"– can immerse users in interacting with them to a greater extent when compared to agents emulating the human form. If so, this can potentially mean that users can feel more engaged while interacting with agents depicted in non-human metaphors or exhibit a higher retention rate due to a greater degree of immersion in their work environment. This, however, remains to be explored, and we empirically address these questions in our work.

Although little research has investigated non-human representations in the context of conversational agents, Nowak and Rauh investigated how people perceive avatars with human, animal, and object forms in terms of their attractiveness and credibility. The study has shown that more anthropomorphic avatars were perceived as more attractive and credible [80]. In addition, several works in games research and electronic commerce have shown that agent avatars with realistic human-like appearances, but with animal features are perceived as being less attractive [31, 76, 90]. This result is either explained by the "uncanny valley effect" [75] or by the mental schema theory as human-like animals elicit categorization tension, which reflects in lower attractiveness or negative attitude towards the avatar [31]. However, other work has shown that older adults have low telepresence in anthropomorphic avatars in human form while they showed higher attractiveness toward animal avatars [18]. This contrasts the argument that iconic images might induce a stronger sense of immersion. Based on these results, one can argue that anthropomorphized non-human avatars might not fit into the human conceptual understanding, resulting in lower user satisfaction. However, discussion over non-human metaphors in existing literature has not been systematically grounded, and the corresponding exploration has been inconsistent.

Some studies in HCI explored techno-spirituality in the form of design fiction, which also can be called transcendent experiences (TXs). For example, Blythe and Buie presented a few imaginary design fictions as a critical design to the research community to explore possible user reactions to techno-spirituality [11]. Buie further made a game prototype

to facilitate the techno-spiritual design that attempts to connect user experience and spiritual experience [15]. Dove and Fayard explored utilizing the *'technology as monster'* metaphor in an early-stage generative design workshop for designers to probe and frame machine learning technology [23]. These studies are relevant to the highest category of '*God*' in the Great Chain of Being that we leverage in this paper.

In summary, we currently lack a clear understanding of how different levels of non-human metaphors can affect user interaction with conversational agents and the potential consequences of using non-human metaphors on user engagement and trust in the agents.

2.3 Conversational Crowdsourcing

Conversational agents have been extensively applied in microtask crowdsourcing. Researchers have used crowdsourcing to aid conversational agents in answering questions [49, 50]. Others have proposed the use of conversational agents to train non-expert crowd workers on picking up domain-specific skills [4]. Crowd-powered conversational systems have also been proposed to overcome challenges that go beyond existing AI technologies [2, 3]. More recently, conversational agents have been introduced in crowdsourcing marketplaces as an interface to interact with crowd workers [69].

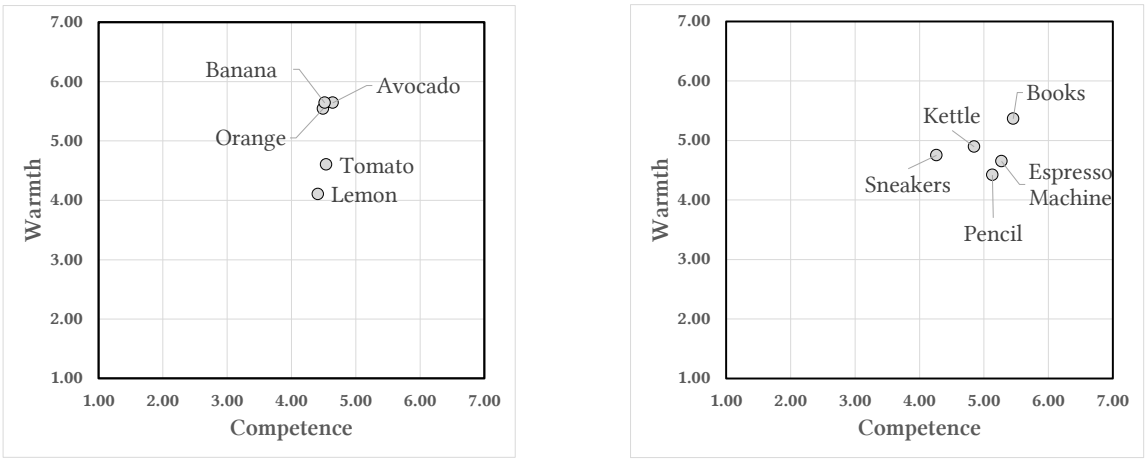
Crowdsourcing work often requires workers to conduct manual human intelligence tasks (HITs), often deployed with similar tasks in large batches [5, 22]. Such repetitive and monotonous tasks risk reducing well-being, creating lower-quality output, and decreasing worker retention [60]. In order to improve worker retention and work quality, Dai *et al.* investigated using "micro-diversions," a small amount of entertainment in between tasks [19].

In addition, prior work used conversational agents to acquire knowledge from crowd work to construct a knowledge base [13]. Recent studies have also developed a conversational interface to support microtask execution in crowdsourcing marketplaces [69], where workers were redirected to Telegram and completed the microtasks with an agent. Consequently, Qiu *et al.* investigated the effect of conversational interface on worker engagement based on different conversational styles and moods. The study showed that conversational interface improved the perceived worker engagement and significantly higher worker retention [86, 87]. Further investigating worker engagement, authors adopted findings from the games research domain and implemented worker avatars to promote self-identification and enhance intrinsic motivation [84]. Results showed that using worker avatars effectively reduced cognitive workload and increased worker retention.

Although prior work has explored avatar customization in crowdsourcing, there is a lack of investigation into how different agent metaphors' visual representations (agent avatars) impact worker engagement, perceived cognitive workload, and trust. In this work, we design a between-subjects study to comprehensively understand the effects of using different metaphor representations on conversational crowdsourcing.

3 STUDY DESIGN

This study aims is to understand how different metaphorical representations could affect workers' subjective perceptions. Therefore, we conducted a 6×2 between-subject experiment to investigate the effect of metaphorical representation of conversational agents on worker engagement, perceived cognitive load, enjoyment, and trust based on the crowdsourcing platform Prolific. Workers were asked to converse with a chatbot with one of six agent representations consisting of five metaphors derived from the Great Chain of Being — `God`, `Human`, `Animal`, `Plant`, `Inorganic Object`, and a `Control` condition with no representation (cf. Figure 4). Workers were also randomly assigned to one of two distinct tasks — image classification tasks and information finding tasks (cf. Figure 5), resulting in 12 experimental conditions [30].



(a) Plant metaphors (b) Object metaphors

Fig. 3. Scatter plots showing the average warmth and competence for the measured metaphors for Inorganic Object and Plant.

3.1 Preliminary Study: Metaphor Sampling and Selection

Previous studies that investigated the impact of users' perception on AI systems reported a clear difference in how users react to the system based on perceived warmth and competency of the system [37, 56]. Therefore, instead of using random metaphors, we chose metaphors that manifest similar degrees of warmth and competency. High warmth and high competency were reported to ensure a positive attitude toward the system [37, 56]. Therefore, being aware of the trade-off that high-competency might elicit lower intention to adapt to the system [35, 37, 56], we decided to unify the warmth and competency metaphors across conditions.

Therefore in this study, we used five different metaphors for each variable from the five levels of the Great Chain of Being apart from the Control condition. As a previous study from Khadpe *et al.* [56] drew upon, the Stereotype Content Model (SCM) [27] was used to sample metaphors coherently. In addition, we decided to use results from previous literature on human and animal metaphors that were reported to correspond to high-warmth and high-competence. As a result, we chose "dog" as a metaphor to represent an Animal agent [93] and "trained professional" as a metaphor to represent a Human agent [56]. We decided not to sample a specific metaphor for God but to represent it plainly as "God" since different representations of God can have socio-cultural and religious connotations that might inadvertently affect participants' perception and behavior [105].

To decide the Plant and Inorganic Object metaphor with high-warmth and high-competence, we conducted two pre-tests on Plant and Inorganic Object on Prolific.[6] We recruited workers with an approval rate of over 90% and who speak English as their native language to ensure quality results. Based on prior literature on human perception of warmth and competence [8], the authors of this paper (who were the experts in the related field) selected a sample of 5 candidate metaphors for each category (Plant and Inorganic Objects) that were deemed to exhibit high warmth and high competence. We chose avocado, banana, lemon, tomato, and orange for the Plant metaphor. Moreover, we chose book, espresso machine, kettle, pencil, and sneakers for the Object metaphor.

[6]https://www.prolific.co

Next, we deployed a study on a crowdsourcing platform in which we asked crowd workers to rate the perceived warmth and competence of 5 candidate metaphors for each category (`Plant` and `Inorganic Objects`, respectively) using 7-point Likert scales. Workers were asked to respond to 22 questions (10 questions for each category and 2 attention check questions to ensure reliability). All items (including the attention check questions) were randomized in the questionnaire. We paid workers on average an hourly wage of GBP £12.19. Finally, we collected 100 responses for both categories (`Plant` and `Inorganic Object`) and filtered out responses from workers who failed to pass at least one of the two attention check questions. This left us with 97 valid responses corresponding to the `Inorganic Object` category and 98 valid responses for the `Plant` category.

As shown in Figure 3, our results demonstrate that the chosen metaphors were in the high-warmth and high-competence categories on average. Based on these findings, we chose Avocado to represent the `Plant` metaphor (cf. Table 1) and Book to represent the `Inorganic Object` metaphor (cf. Table 2). These were reported to exhibit the highest perceived warmth and perceived competence among the subjects considered in the preliminary study. Table 3 shows our final selection of metaphors to represent each Great Chain of Being category.

Table 1. Warmth and competence values (average ± standard deviation) corresponding to `Plants` in the preliminary study.

	Avocado	**Tomato**	**Banana**	**Orange**	**Lemon**
Competence	**4.64 ± 1.63**	4.54 ± 1.60	4.52 ± 1.67	4.49 ± 1.58	4.41 ± 1.61
Warmth	**5.64 ± 1.44**	4.60 ± 1.72	5.64 ± 1.34	5.54 ± 1.58	4.10 ± 1.90

Table 2. Warmth and competence values (average ± standard deviation) corresponding to `Inorganic Objects` in the preliminary study.

	Books	**Sneakers**	**Kettle**	**Espresso machine**	**Pencil**
Competence	**5.45 ± 1.58**	4.26 ± 1.80	4.85 ± 1.62	5.27 ± 1.56	5.13 ± 1.55
Warmth	**5.36 ± 1.41**	4.75 ± 1.66	4.90 ± 1.64	4.65 ± 1.68	4.42 ± 1.70

Table 3. Six experimental conditions to represent different metaphors based on the Great Chain of Being (informed by the preliminary study), including a Control condition to help analyse the impact of these metaphorical representations of the agent.

#	**Category**	**Selected Metaphor**
1	God	God
2	Human	Trained Professional
3	Animal	Dog
4	Plant	Avocado
5	Inorganic object	Book
6	Control	No representation

Although the SCM framework has been highly influential and has been repeatedly adopted in previous literature, it has received criticism for inconsistent operationalization of both SCM dimensions: two factors of warmth and competence [12, 43]. To reduce such concern, we used the recent study conducted by Halkias and Diamantopoulos that

suggested a more accurate and consistent operationalization of measuring warmth and competence [43]. In addition, one might be concerned about using the SCM on subjects like plants or animals. Nevertheless, previous studies, notably from the consumer psychology domain, have demonstrated that the perceived warmth and competence of non-human entities can be reliably measured using a Likert scale [1, 38].

3.2 Microtask Design

We chose the two task types of Information Finding and Image Classification, covering two data types (text and images). These task types are prevalent in microtask crowdsourcing marketplaces [22, 30] and have been the subject of recent research. Participants were assigned to a single experimental condition on one of these task types, with one of the six different agent metaphors. Therefore, our study resulted in 12 experimental conditions (2 task types $\times$ 6 agent metaphors). Previous studies in conversational microtask crowdsourcing employed agents and designed conversations that had no direct relation to the task at hand (e.g., Information Finding task of finding restaurants, and Image Classification task with classifying animal species) [86]. In this work, however, we argue that a metaphorical understanding can be further acquired by demonstrating the characteristics and traits of the metaphor through explicit conversation and not only limiting this to a visual representation through an agent avatar, for example, by explicitly conveying that avocado is nutritious, and that a book relates to knowledge through a well-defined narrative. This design choice ensures that users conversing with the conversational agent can fully experience the metaphor of the agent. Therefore, we used the same topic, a task related to a bird, in both task conditions to unify the narrative between each agent and the task that users are asked to execute. Workers from both task conditions, either Image Classification or Information Finding tasks, are required to look and think about birds and listen to the same narrative from a metaphor agent that they are talking to (e.g., "*God created different shapes of bird, and the God is asking you to either find information about birds or classify images about birds.*").

We decided to use Caltech-UCSD Birds 200 Dataset [100], an image dataset comprising 200 bird species with annotations. Each task comprised a total of 36 questions about 36 different birds. Participants could answer as many questions as they wished to, and they were free to leave whenever they wanted to, after completing of 10 mandatory tasks.

(1) *Information Finding (IF) Task.* Workers were asked to find either an Order, a Family, a Genus, or a Species of a given bird on Wikipedia (cf. Figure 5 (c),(d)). These information finding tasks are based on the taxonomic classification system of birds. The Class is divided into Orders; the Orders are composed of Families; the Families are divided into Genera, and the Genera are composed of distinct Species.

(2) *Image Classification (IC) Task.* Workers were asked to analyze images of birds and to classify the shape of their beak, also called 'a bill' (cf. Figure 5 (a) and (b)). This task was adapted from a previous study by Wah C. *et al.* [100]. We added images of 10 birds that were not included in the Caltech-UCSD Birds 200 dataset for this task to balance the number of birds across nine distinct categories of beaks sampled for the set of 36 tasks, a criterion that was difficult to fulfill with the existing dataset. The IC task was chosen to ensure that different data types were tested in the study — text for IF tasks and images for IC tasks — because previous studies revealed different UES-SF results between image-based tasks and text-based tasks [86]. In addition, IC tasks are one of the most common task types prevalent on crowdsourcing platforms. Since conversational interfaces that are HTML-based (e.g., TickTalkTurk [88]) can be easily ported from traditional web interfaces, it will be interesting to see the impact of the IC tasks presented in a conversational style.

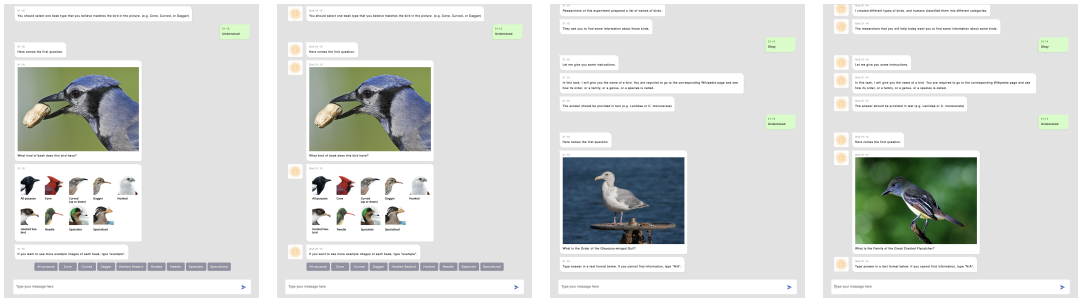
3.3 Agent Avatar Design

As shown in Table 3, we selected six metaphors to represent the conversational agent. We decided not to visualize the Control condition to avoid any visual bias that might affect the participant's impression of the agent. In addition, previous research conducted within conversational crowdsourcing has tested interfaces without agent metaphors, and their result can serve as a baseline to interpret our treatment conditions [84, 86]. Next, we designed five avatars to represent the different metaphors (Figure 4). The Naturalism-stylization framework is one that designers can decide on in the virtual agent visualization [40, 42]. We used stylized visuals for our agent because they have been reported to affect users' interactions positively. Stylized pedagogical agents are more likely to be chosen by female students [42]. Also, stylized e-commerce agents are reported to produce a higher social perception of an agent, perceived website social presence, and perceived website social support compared to an agent with naturalistic visualization [97].

(a) God (b) Human (c) Animal (d) Plant (e) Inorganic Object

Fig. 4. Conversational agent avatars to visually represent the five different metaphors.

3.4 Worker Interface



(a) Worker interface (IC) ***without*** agent avatar

(b) Worker interface (IC) ***with*** agent avatar (god)

(c) Worker interface (IF) ***without*** agent avatar

(d) Worker interface (IF) ***with*** agent avatar (god)

Fig. 5. The comparison of conversational interfaces with and without agent avatar, and between two microtask types.

The conversational interface was built using TickTalkTurk [88], a Web-based application that is created with HTML, CSS, and Javascript. For Control conditions, no avatars were displayed to prevent bias from participants (Figure 5 (a) and (c)). The interface displayed the agent avatar next to the text bubble of the agent's dialogue with five other metaphors.

Workers could answer microtask questions in two different input types: free text and multiple-choice. Free text input required workers to answer questions by typing and sending them to the conversational agent as a chat interface

message. Multiple choice input allows workers to choose one answer using customized radio buttons. We employed multiple choice options for IC tasks and free text for IF tasks. After the microtasks, workers were asked to move to a Google Forms page to complete post-task surveys.

3.5 Conversation Design

All the metaphor conditions shared the same structure and order in dialogues. However, different words and vocabularies are used to enhance the experience of workers interacting with the metaphor. Table 4 highlights dialogue excerpts that showcase the variation in conversation templates between the metaphors. The complete conversation template can be found on our companion web page.[7]

Table 4. Conversation template for the metaphor conditions, with excerpts demonstrating variations in narratives across the conditions. The complete conversational template is available at our companion web page: https://sites.google.com/view/agent-metaphors.

	Opening greetings	**Agent Narrative**
God	Greetings, I am the lord thy god. You shall help researchers at a university by participating in this research. So I ask you, does this sound good to you?	I created different types of birds, and they all have different shapes of beaks.
Human	Hello, pleasure to meet you! I'm a trained professional who will guide you here. Researchers in a university asked me to facilitate this research. Does this sound good to you?	They (researchers) asked me to help you because I am trained for helping people like you!
Animal	Woof woof, nice to meet you! I'm a dog that will help you conduct this research (…)	They asked me to help you because I am super good at finding stuff!
Plant	Hello, nice to meet you! I'm an avocado that will help you conduct this research (…)	They asked me to help you because I'm full of healthy nutrition and I can nourish you throughout this task.
Inorganic Object	Hello, nice to meet you! I'm a book that will help you conduct this research (…)	They asked me to help you because I'm full of information and knowledge.
Control	Hello! Can you help researchers in a university by conducting this research?	N/A

While designing the conversation, we ensured that the agent acted as a "moderator" of the task to make the narrative consistent across all metaphors. Therefore the agent asked workers if they would like to help the researchers rather than the agents themselves. If agents were not a moderator and intended to ask specific tasks to a worker (e.g., that a dog wants to hunt or that a book is missing information), it might affect participants' perception of the task and bias their responses.

However, we still had to embody an experiential aspect of a specific metaphor. Therefore, we explained why the agent became a moderator by raising their characteristics and attributes. After they received the task instructions, this narrative was presented to workers (Table 4).

3.6 Measures

The dependent variables of our experiment are worker engagement, task load, enjoyment, and trust.

Worker Engagement is measured using a short form of the User Engagement Scale (UES-SF) [81], which is a scale that measures self-reported user engagement.

[7] https://sites.google.com/view/agent-metaphors

Enjoyment is measured using the Intrinsic Motivation Inventory (IMI) [71], an instrument that measures participants' subjective experience on the target activity. IMI has been widely adopted in research and has been validated in different contexts [66, 70]. We adopted a relevant subset of the IMI to reduce the number of post-task questions that we asked workers to respond to. We covered two dimensions of the IMI, spanning 10 questions that were most relevant to our research: Interest-Enjoyment (INT-ENJ) and Perceived Competence.

Cognitive Task Load is measured with the NASA-TLX questionnaire [47, 48]. Through this measure, we investigate any significant difference in perceived cognitive load between agent metaphors.

Trust is measured through the Trust in Automation (TiA) questionnaire [54, 59]. As we did with the IMI questionnaire, we used a relevant subset of TiA to limit the post-task questions that workers were asked to complete. We selected *Propensity to Trust* and the *Trust in Automation* dimensions, which spanned 5 questions in total. TiA is a validated questionnaire [59] that has been widely used to measure trust in human-AI interaction [98, 102].

Moreover, other measures such as task execution time and output accuracy are also recorded for evaluating worker performance and behavior.

3.7 Participants

We recruited workers from the Prolific crowdsourcing platform. To avoid potential biases, workers who participated in the preliminary study were not allowed to partake in this main study. We recruited workers who speak English as their first language and whose approval rates were greater than 93%. Workers were not allowed to accept and execute the task multiple times and were randomly assigned to one of twelve conditions (6 agent representations $\times$ 2 task types). We compensated workers GBP £1.13 (USD $1.6) and GBP £1.88 (USD $2.6) for Image Classification and Information Finding tasks, respectively. Furthermore, we rewarded workers 4 pennies for answering each optional question. As a result, workers received an hourly wage of £6.5 on average (USD $9.0 / hour).

A power analysis using G*Power [26] indicated that a minimum sample size of 27 participants for each group is needed. Therefore, we recruited 360 unique workers to account for potential participant exclusion in our analysis. We excluded 19 workers who failed to correctly answer one or more of the four attention check questions. Therefore, we only consider responses from 341 workers for further analysis.

3.8 Experimental Setup and Procedure

First, workers were redirected to the conversational interface we designed. Here, workers were asked to complete 10 mandatory microtasks. After workers completed the mandatory tasks, the conversational agent notified them that the mandatory session was finished and that they could continue with optional questions. At this stage, a bigger size image file of the agent avatar was displayed on the screen to reinforce the agent metaphor that they were interacting with (Figure 6). This break message is induced after 10 consecutive questions in the Image Classification task and 5 in the Information Finding task. The different interval between task types was based on a task completion time we measured a priori to the experiment. We decided to design this micro-diversion to prevent fatigue and boredom of workers based on prior work [19, 89], inform workers on the number of optional questions left, and remind them that they could stop anytime they wished.

After workers finished completing tasks, they were redirected to the survey page based on Google Form.[8] They were asked to answer the 5-point Likert scale of the User Engagement Scale Short Form (UES-SF), the NASA Task Load Index

[8]https://www.google.com/forms/about/

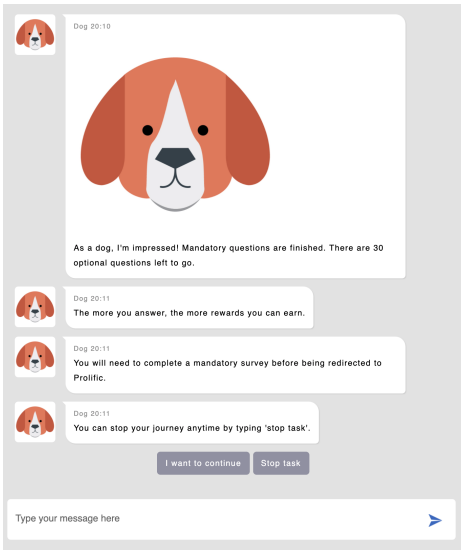


Fig. 6. The conversational agent (`Animal` metaphor) reminding workers that they can stop after 10 mandatory tasks. This message was shown either every 10 tasks (IC) or 5 tasks (IF) based on a priori estimated task completion time for each task type.

form (NASA-TLX), a subscale of Intrinsic Motivation Inventory (IMI) form, and a subscale of Trust in Automation form (TiA). Lastly, workers answered one demographic question about educational background. The ethics committee of our institute has approved this study.

4 RESULTS

4.1 Worker Demographics

Among 341 unique workers, 61.5% were female, and 38.5% were male. 62.9% of workers were under 29 years old, 38.2% between 30 and 49, and 4.1% over 50. In addition, 64.5% of the workers reported that they received higher than Bachelor's degree in educational level.

4.2 Worker Engagement

According to the normality test (Shapiro-Wilk tests, $\alpha = 0.05$), UES-SF scores did not come from a normal distribution. Therefore we conducted the Kruskal-Wallis test to check statistical differences across different conditions. The distributions of UES-SF scores are shown in Figure 7 using boxplots, while the mean values and standard deviations in all the dimensions are reported in Table 5. The detailed statistical report can be found in Appendix A.

A Kruskal-Wallis test revealed that different metaphorical representations significantly affect the overall UES-SF score, the Perceived Usability, and Aesthetic Appeal subscales. We thereby conducted a post-hoc analysis using Mann-Whitney U tests using a Bonferroni-adjusted alpha level of .008 (0.05/6) to compare all pairs of metaphors. For the overall UES-SF score, the `Object` metaphor (Book) received a significantly lower score than the `Animal` metaphor (Dog). Moreover, the Aesthetic Appeal of the Book metaphor was significantly lower than the Dog metaphor.

Afterward, each metaphor was compared with the `Control` condition. As post-hoc Kruskal Wallis tests with Bonferroni corrected alpha values revealed that the Perceived Usability of the `Control` condition was significantly

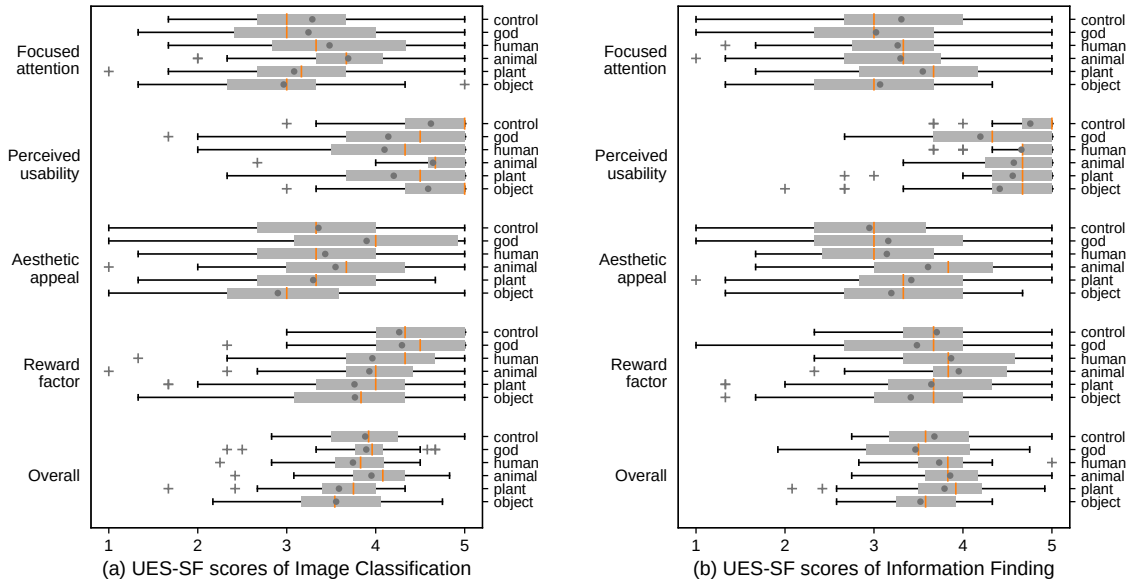


Fig. 7. Boxplots showing worker engagement measured by UES-SF, where red lines represent medians and black points represent mean values.

Table 5. The UES-SF score ($\mu \pm \sigma$: mean and standard deviation) of all task types with six metaphor conditions.

	Control	God	Human	Animal	Plant	Object	Overall
				Image Classification			
Focused Attention	3.29 ± 0.86	3.24 ± 1.15	3.48 ± 1.01	3.69 ± 0.75	3.08 ± 0.85	2.97 ± 0.90	3.29 ± 0.94
Perceived Usability	4.62 ± 0.59	4.14 ± 1.03	4.10 ± 1.02	4.64 ± 0.50	4.20 ± 0.89	4.59 ± 0.58	4.39 ± 0.81
Aesthetic Appeal	3.36 ± 1.09	3.90 ± 1.12	3.43 ± 0.87	3.55 ± 0.90	3.30 ± 0.94	2.90 ± 1.03	3.39 ± 1.03
Reward Factor	4.26 ± 0.67	4.29 ± 0.73	3.96 ± 0.88	3.93 ± 0.87	3.76 ± 0.93	3.77 ± 0.88	3.99 ± 0.85
Overall	3.88 ± 0.54	3.89 ± 0.56	3.74 ± 0.52	3.95 ± 0.54	3.59 ± 0.60	3.56 ± 0.63	3.77 ± 0.58
				Information Finding			
Focused Attention	3.31 ± 1.02	3.02 ± 0.98	3.27 ± 0.89	3.30 ± 1.08	3.55 ± 0.94	3.07 ± 0.84	3.25 ± 0.96
Perceived Usability	4.76 ± 0.42	4.20 ± 0.80	4.66 ± 0.42	4.57 ± 0.49	4.56 ± 0.59	4.41 ± 0.82	4.52 ± 0.63
Aesthetic Appeal	2.95 ± 1.14	3.16 ± 1.15	3.14 ± 0.88	3.61 ± 1.02	3.42 ± 1.00	3.20 ± 0.82	3.25 ± 1.01
Reward Factor	3.71 ± 0.74	3.48 ± 1.15	3.87 ± 0.78	3.95 ± 0.75	3.65 ± 1.04	3.41 ± 0.90	3.68 ± 0.92
Overall	3.68 ± 0.60	3.47 ± 0.73	3.73 ± 0.49	3.86 ± 0.53	3.79 ± 0.71	3.52 ± 0.53	3.68 ± 0.61

higher than the God metaphor. In addition, we found that the Perceived Usability of the `Control` condition without any agent metaphors was significantly higher than the agents with metaphorical representations.

We exploratively looked at the difference in UES-SF score between two task types. Mann-Whitney test revealed a significant difference in the Reward Factor dimension. Reward Factor in Image Classification task ($Md = 4.00$, $n = 168$) was significantly higher than the Information Finding task ($Md = 3.67$, $n = 173$), $U = 11336.00$, $z = -3.543$, $p < .001$, with small effect size $r = .19$.

Summary: `Animal` (Dog) received a significantly higher overall UES-SF and Aesthetic Appeal subscale score than the `Object` (Book), while no significant difference was found in the Perceived Usability subscale. Perceived Usability for `God` was significantly lower than the `Control` condition.

4.3 Intrinsic Motivation

According to the normality test (Shapiro-Wilk tests, $\alpha = 0.05$), IMI scores did not come from a normal distribution. We, therefore, conducted the Kruskal-Wallis test to find statistical differences across different conditions. Worker IMI scores are shown in Figure 8 with boxplots, and the mean values and the standard deviation are reported in Table 6. The detailed statistical report can be found in Appendix B.

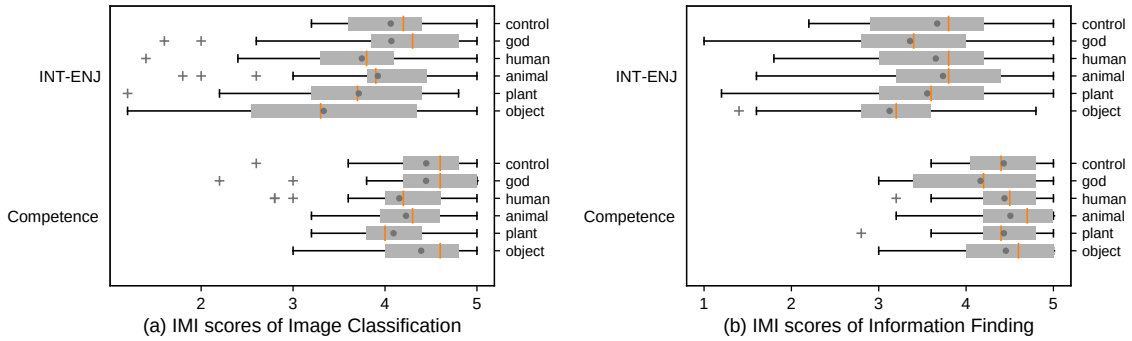


Fig. 8. Boxplots showing worker intrinsic motivation measured by IMI, where red lines represent medians and black points represent mean values.

Table 6. The IMI score ($\mu \pm \sigma$: mean and standard deviation, unit: the number of optional question answered) of all task types with six metaphor conditions.

	Control	God	Human	Animal	Plant	Object	Overall
				Image Classification			
Interest-Enjoyment	4.06 ± 0.55	4.07 ± 0.97	3.75 ± 0.79	3.92 ± 0.83	3.71 ± 0.81	3.33 ± 1.16	3.80 ± 0.90
Competence	4.45 ± 0.56	4.45 ± 0.66	4.16 ± 0.62	4.23 ± 0.50	4.09 ± 0.51	4.39 ± 0.55	4.30 ± 0.58
				Information Finding			
Interest-Enjoyment	3.67 ± 0.90	3.36 ± 1.02	3.65 ± 0.82	3.74 ± 0.87	3.55 ± 0.97	3.12 ± 0.77	3.51 ± 0.91
Competence	4.43 ± 0.45	4.17 ± 0.67	4.44 ± 0.50	4.51 ± 0.56	4.43 ± 0.50	4.46 ± 0.59	4.40 ± 0.55

After conducting a Kruskal-Wallis test, we found a statistically significant difference in the Interest-Enjoyment dimension between at least two metaphor groups. Therefore, the Man-Whitney tests with a Bonferroni-adjusted alpha level of .008 (0.05/6) were conducted between different conditions. We consistently found that the `Object` (Book) metaphor received a lower Interest-Enjoyment score than any other metaphor. Especially, Interest-Enjoyment for the `Object` (Book) was significantly lower than the `Human` (Trained Professional), `Animal` (Dog), and the `Control` condition.

Although not statistically significant, `Object` (Book) still received a notably lower score than the `God` and the `Plant` (Avocado) metaphors after the Bonferroni correction.

According to the Mann-Whitney test, there was a significant difference in Interest-Enjoyment dimension between two different task types. Image Classification ($Md = 4.00$, $n = 168$) scored higher than the Information Finding ($Md = 3.60$, $n = 174$), $U = 11508.00$, $z = -3.331$, $p = .001$, with a small effect size $r = .18$.

> ***Summary***: Overall, the `Inorganic Object metaphor` (Book) relates to lower interest-enjoyment than other conditions. In addition, the Image Classification task received a higher interest-enjoyment score than the Information Finding task.

4.4 Perceived Cognitive Task Load

According to the normality test (Shapiro-Wilk tests, $\alpha = 0.05$), TLX scores did not come from a normal distribution. We, therefore, conducted the Kruskal-Wallis test to find statistical differences across different conditions. The distributions of overall TLX scores are shown in Figure 9. The mean values and standard deviations of TLX scores in each dimension are reported in Table 7. The detailed statistical report can be found in Appendix C.

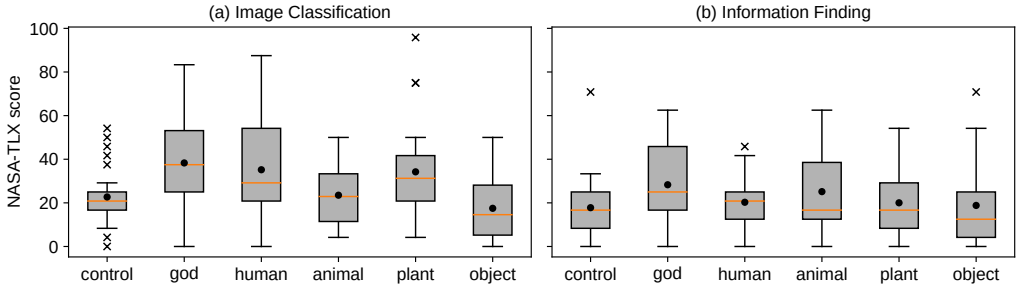


Fig. 9. Boxplots showing perceived cognitive task load measured by NASA-TLX, where red lines represent medians and black points represent mean values.

Table 7. The NASA-TLX score ($\mu \pm \sigma$: mean and standard deviation, unit: the number of optional question answered) of all task types with six metaphor conditions.

	Control	God	Human	Animal	Plant	Object	Overall
Image Classification	22.70 ± 12.82	38.30 ± 20.45	35.19 ± 23.35	23.51 ± 13.28	34.23 ± 20.71	17.50 ± 14.70	28.25 ± 19.23
Information Finding	17.79 ± 14.79	28.30 ± 20.15	20.28 ± 11.77	25.15 ± 17.77	20.03 ± 14.65	18.82 ± 17.35	21.75 ± 16.46

A Kruskal-Wallis test revealed a statistically significant difference between at least two metaphor groups concerning the overall TLX score. Therefore, we conducted post-hoc Mann-Whitney tests using Bonferroni-adjusted alpha level of .008 (0.05/6). Consistent with the IMI result, the `Object` (Book) metaphor resulted in a lower perceived task load in comparison to the `God`, `Human` (Trained Professional), and `Plant` (Avocado) metaphors with a statistically significant difference. The cognitive load for the `Control` condition was significantly lower than the `God` metaphor.

According to the Mann-Whitney test, image classification task received a significantly higher score than the information finding task in overall TLX score , IC: $Md = 25.00$, $n = 168$ | IF: $Md = 16.67$, $n = 173$ | $U = 11462.50$, $z = -3.384$, $p = .001$, small effect $r = .18$.

> ***Summary***: Results show that the `Inorganic Object` metaphor (Book) corresponds to a lower cognitive load. The perceived cognitive load corresponding to the God metaphor was significantly higher than the `Control` condition.

4.5 Trust

Trust score was not normally distributed (Shapiro-Wilk tests, $\alpha = 0.05$). Therefore we conducted the Kruskal-Wallis test and found no significant difference in dimensions of Trust in Automation (TiA) scales across different metaphor conditions and between task types (Table 18). Also, no difference was found between Control conditions and conditions with metaphor representation. Table 8 shows the unweighted TiA score of the two dimensions we used, and Figure 10 shows the TiA score across six metaphor conditions in two task types. The detailed statistical report can be found in Appendix D.

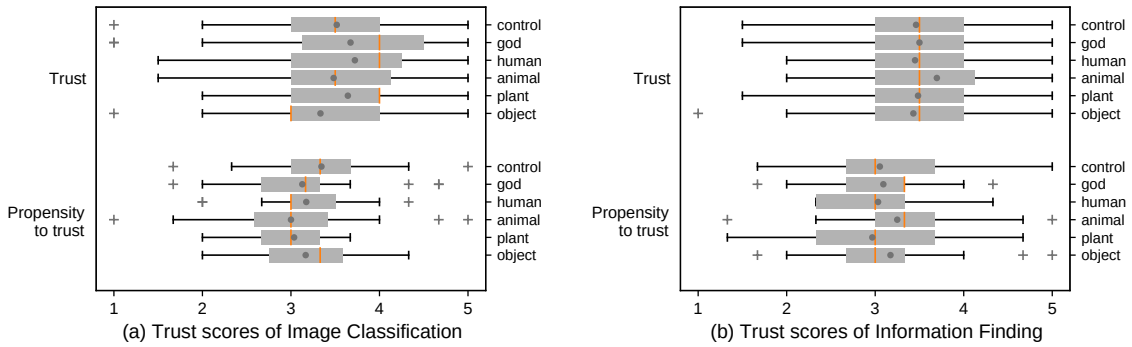


Fig. 10. Boxplots showing worker trust measured by Trust in Automation, where red lines represent medians and black points represent mean values.

Table 8. The TiA dimension score ($\mu \pm \sigma$: mean and standard deviation, unit: the number of optional question answered) of all task types with six metaphor conditions.

Metaphors	**Control**	**God**	**Human**	**Animal**	**Plant**	**Object**	**Overall**
				Image Classification			
Trust	3.52 ± 1.02	3.67 ± 1.16	3.72 ± 0.92	3.48 ± 1.00	3.64 ± 0.78	3.33 ± 0.93	3.56 ± 0.97
Propensity to Trust	3.34 ± 0.69	3.13 ± 0.74	3.17 ± 0.55	3.00 ± 0.90	3.04 ± 0.45	3.17 ± 0.55	3.14 ± 0.66
				Information Finding			
Trust	3.46 ± 0.85	3.50 ± 0.93	3.45 ± 0.75	3.70 ± 0.82	3.48 ± 0.83	3.43 ± 0.90	3.50 ± 0.84
Propensity to Trust	3.05 ± 0.79	3.09 ± 0.62	3.03 ± 0.59	3.25 ± 0.76	2.97 ± 0.84	3.17 ± 0.74	3.09 ± 0.72

Summary: We found no significant differences in trust across the different metaphorical representations and task types.

4.6 Output Accuracy and Task Execution Time

In this study, we also measured output accuracy and active task execution time to understand worker performance. According to the normality test (Shapiro-Wilk tests, $\alpha = 0.05$), results of worker performance did not come from normal distributions. The distributions of worker performance measures are shown in Figure 11. We found no significant result in output accuracy and task execution time between different metaphors. The detailed statistical report can be found in Appendix E.

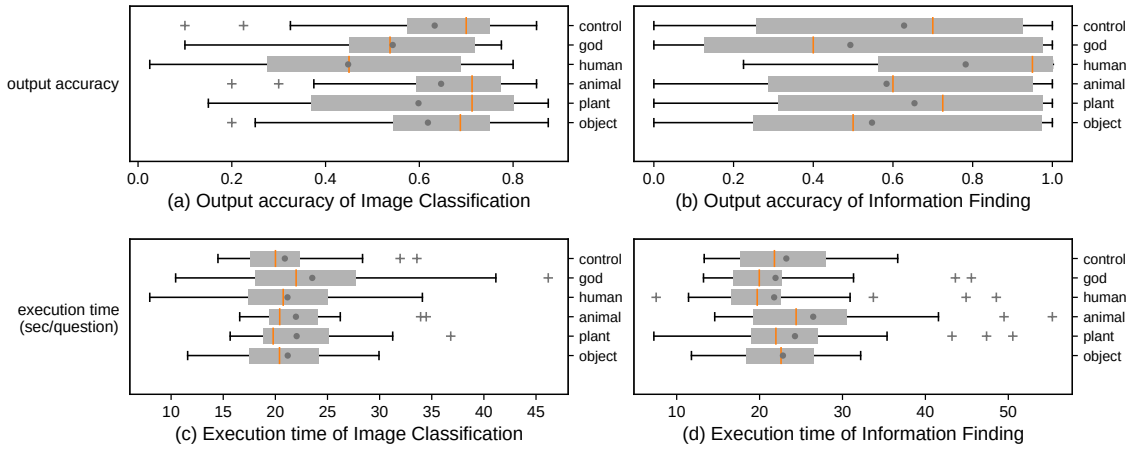


Fig. 11. Boxplots showing workers' output accuracy and active task execution time, where red lines represent medians and black points represent mean values.

Summary: We found no statistically significant difference in output accuracy and task execution time across different metaphorical representations, control conditions, and task types.

Table 9. A table with a summarized view of significant results. Results that are considered more positive are colored in light blue, while negative results are colored in light red. Trust and Accuracy measurements were omitted since they did not vary significantly across the experimental conditions.

	User Engagement Scale	Perceived Usability (UES)	Aesthetic Appeal (UES)	Interest-Enjoyment (IMI)	Cognitive Load
God	-	↓ than Control	-	-	↑ than Object, Control
Human	-	-	-	↑ than Object	↑ than Object
Animal	↑ than Object	-	↑ than Object	↑ than Object	-
Plant	-	-	-	↑ than Object	↑ than Object
Object	↓ than Animal	-	↓ than Animal	↓ than Human, Animal, Control	↓ than God, Human, Plant
Control	-	↑ than God	-	-	↓ than God

5 DISCUSSION

Metaphorical understanding of agents has been identified as a potential explanation for how users expect a system to behave and how users react to the system [56]. While discussion within the HCI community has primarily focused on human and anthropomorphic representations, agents with non-human representation are abundantly used in the real world despite a lack of complete understanding of their effects. Although such a tendency to apply non-human metaphors to agents can be explained by the desire to create a particular impression of the system, few works have explored the impact of non-human metaphor agents. This lack of attention is partly due to anthropomorphism shown to affect and enhance user interest, engagement, and credibility towards the agent [58, 80, 102, 103]. However, some studies have shown that anthropomorphic images are perceived as less credible and likable [78]. Moreover, when the degree of anthropomorphism in an image inversely matched the system's ability, people's perception of the system deteriorated [35].

Although few studies investigated the impact of a virtual agent with animal and inorganic representations [18, 31, 76, 80], they have not been grounded in the Conceptual Metaphor theory but rather within the context of manipulating the degree of anthropomorphism. Therefore, a more systematic approach is needed to understand the non-human agent representation based on the Conceptual Metaphor theory. Hence, we borrowed the Great Chain of Being framework to understand the effect of a range of hierarchical cross-species metaphors. In this paper, we carried out an experimental study to understand the impact of different agent metaphors based on the Great Chain of Being framework in the context of conversational crowdsourcing.

Our findings suggest the absence of a consistent trend across different agent metaphors regarding their impact on the considered dependent variables and the task types. Nonetheless, the `Object` (Book) metaphor generated the most significant difference with other conditions. For example, the `Animal` (Dog) metaphor resulted in a higher worker engagement and a higher cognitive task load than the `Object` metaphor. In contrast, the `Inorganic Object` (Book) metaphor, which is the lowest level in the hierarchical Great Chain of Being framework, stood out in its effect on lowering the perceived cognitive load of the workers, yet impairing workers' interest and enjoyment while conducting the tasks. Similarly, the Image Classification task resulted in a higher Reward Factor (UES) and Interest-Enjoyment scale (IMI) than the Information Finding task, but cost more cognitive task load. This contrasts previous findings that when workers are more engaged with work, they perceive less cognitive load [84, 106]. We could potentially explain this variance that the `Animal` (Dog) metaphor dragged workers' attention more towards the agent than the task at hand, and cost workers' working memory. Exploring the exact mechanism behind it is a promising research opportunity.

Contrary to our understanding informed by previous work [58, 80, 102, 103] that the `Human` (Trained Professional) metaphor will enhance engagement, intrinsic motivation, and trust in the context of crowd work, our result shows that the `Human` metaphor had no significant difference over non-human metaphors. The `Human` metaphor showed significantly higher Interest-Enjoyment than the `Object` metaphor, but its effect was smaller than the `Animal` and `Control` condition. In addition, it resulted in a higher cognitive load than the `Object` metaphor with a statistically significant difference. This result can be potentially explained by the findings of Garau *et al.*, where authors showed that when agent representation does not meet users' expectations of a system, this hampers user perception of the system. Since our task design used a relatively simple conversational agent with limited ability, this may have caused workers to be disappointed with the system's actual ability [35].

The `God` metaphor, the highest level in the hierarchical Great Chain of Being framework, consistently resulted in a higher cognitive task load than the `Control` condition and the lowest level in the framework — `Inorganic Object`

(book). Also, the God metaphor resulted in a lower Perceived Usability than the Control condition without an agent metaphor. This can be related to the previous finding that supports contrast theory [95]; users' higher expectations of the system capability can lead to disappointment when said expectations are not met, leading to lower satisfaction [67]. Users are less forgiving of lower performance and more willing to adapt to the system when interacting with metaphors of a higher competence [56]. Moreover, workers' lower engagement and higher cognitive task load observed in the condition with the God metaphor can also be explained via the lens of worker autonomy. Previous studies argued that the autonomous nature of crowd work is a prime motivator for crowd workers [16, 53]. Therefore, the God metaphor instructing workers to do tasks may have appeared to violate their agency, which could have lowered their engagement and thereby increased the perceived cognitive load. In addition, the God metaphor impairing worker experience can also be explained that the God may exhibit high competency but arguably conveys low warmth in our experiment setting.

In this study, we found no significant difference in trust across the different metaphorical representations of the conversational agent. This contrasts a previous study conducted in the domain of explainable artificial intelligence (XAI) within the speech recognition context, where an embodied agent (in human form) that was used to explain system behavior resulted in a higher user trust in the system [102]. However, this could be possibly explained by the fact that the study mentioned above was conducted with a multi-modal human agent in a face-to-face interaction context, while our study only represented the agent in a graphical avatar with textual interaction.

We found that some of our experimental conditions with different metaphorical agent representations showed no significant improvement in interest and enjoyment, trust, and cognitive load compared to the Control condition. This contrasts with a previous finding from conversational crowdsourcing research that allowing workers to customize their avatars while conversing with a Human agent led to a higher worker engagement and lower cognitive load [84]. Our result shows that worker engagement is only improved when workers deploy self-identified avatars in crowdsourcing, while the agent they interact with does not affect. We wonder if this result was due to the lack of autonomy that workers could manifest as the mentioned work allowed workers to customize their avatars. Future work can explore the potential of facilitating the customization of agent metaphors or aligning agent metaphors according to worker preferences.

5.1 Design Implications for HCI

Using an agent metaphor that locates higher than Human in the Great Chain of Being framework may disappoint workers in the system's ability due to high expectations, leading to a lower engagement and a higher perceived cognitive task load. We found that the Human metaphor in the crowdsourcing context did not show overpowering benefits compared to the Animal (Dog) metaphor or a Control condition without any representation in enhancing engagement and intrinsic motivation, and Inorganic Object (Book) metaphor in lowering the cognitive load. We found clear evidence suggesting that the God metaphor should be avoided in conversational crowdsourcing agent design. The conversational agent represented using the God metaphor was perceived to have significantly lower usability than the Control condition while corresponding to a higher cognitive load than the Control condition and the Object (Book) metaphor. Task designers should be mindful of these trade-offs while making design choices about agent representation in conversational crowdsourcing.

The Inorganic Object (Book) metaphor was found to reduce the cognitive load while decreasing the interest and enjoyment of the context. Therefore, in tasks with higher complexity [107], task requesters or designers can consider using an Inorganic Object metaphor to represent the conversational agent. On the contrary, when the

task is relatively easy but requires more interest due to its less challenging work, one may consider using the Animal metaphor to increase the interest and enjoyment during task completion.

Lastly, in a conversational crowdsourcing context, an agent without any metaphorical representation can be a safe choice to ensure higher perceived usability than the God metaphor and a higher interest-enjoyment than the Object metaphor, while not costing a significant difference in cognitive task load than other agents with GCoB metaphors.

5.2 Caveats, Limitations and Future Work

One of the goals of this research was to improve worker engagement. Since it is well-understood that a primary motivation for workers is typically to maximize their monetary rewards in crowdsourcing marketplaces, it is not easy to distill the effect of different metaphorical representations on worker engagement. However, our design choices and controlled experiment allow us to draw meaningful conclusions. Pay per unit time was identical across all the conditions and the task types.

Our experimental design considered an Image Classification task to measure the impact of different representations of CAs across two different data types (text and image). Prior work has discussed how different types of tasks could be adapted to the conversational crowdsourcing setting [69] to improve worker engagement. This entails a trade-off between the effort required for task adaptation with potential gains through improved worker engagement. However, we acknowledge that the Image Classification task may not be intuitively conversational. Future work can use tasks that are inherently more flexible from the design standpoint and arguably benefit more from a conversational interface (for example, image annotation tasks).

Metaphors can be interpreted, comprehended, and received differently by workers based on their cultural background (e.g., In the U.S., modern business is understood metaphorically through American football, and in Japan, their national value is understood through the Japanese garden [32, 33]). Although we controlled the demographic pool of participants in our study and restricted participation to those workers who could speak English as a native language (e.g., US, Canada, UK), it is possible that workers interpreted metaphorical representations differently (i.e., we can expect there to be individual differences among workers on their perception of agent metaphors).

In addition, the only agent that was not gender-neutral corresponded to the Human metaphor (Figure 4 (b)) agent avatar. This design choice was based on a previous study with an agent embodying female face enhanced participants' rapport, perceived human-likeness, and trustworthiness towards the agent [94]. However, certain genders might affect a user's perception and reaction to the agent. We aim to address this limitation in our imminent future work. One way to address this could be by adopting human metaphor agents from different genders to see if there is any difference in worker engagement based on the agent's gender.

We intentionally chose the general metaphor 'God' rather than representing a specific God to avoid implying any social or religious connotation. By making this design choice, we did not explicitly control for this particular agent metaphor's perceived competence and warmth. As Gilad *et al.* have shown, warmth perception of an AI system can overpower the perceived competence of the system [37]. Future work can explore how different metaphors at the level of 'God,' which are perceived to have high competence and high warmth, would affect work dynamics in conversational crowdsourcing.

Furthermore, we acknowledge the overhead of designing and adapting metaphorical representations of agents in the crowdsourcing context. However, the HCI community can build adaptable toolkits for agent representations to better understand how metaphorical representations of agents can influence work in conversational crowdsourcing. From a requesters' standpoint, in the absence of such solutions, one can argue that such overhead can be seen as a reasonable

trade-off since an appropriate metaphor design for conversational agents could benefit large batches of crowdsourcing microtasks. After all, prior works that have explored mechanisms to improve worker engagement in crowdsourcing have argued that meaningful benefits can be reaped through said mechanisms by retaining workers in large batches of tasks [21, 29, 84].

Finally, it can be argued that the hierarchical categories within the Great Chain of Being are not mutually exclusive and, in some cases, rather complementary. An example can be a metaphor of a humanoid, a human-like robot, which also is one of the most common representations of commercial chatbots. A humanoid has a special place in the Great Chain of Being, as it is a combination of a general human constitution with inorganic objects. As the humanoid metaphor case shows, when non-human agents are used, they are in between the lines of great chains of being as a conversation is a highly human trait. Our study deliberately avoided anthropomorphism to investigate strict mapping on the Great Chain of Being, although complemented by the fact that they spoke the human language, a necessary pre-requisite. Along this vein, this paper has investigated the performance of the metaphorical agent within the Great Chain of Being framework to gain insights into the conceptual structure. However, how the result of this study compares with "traditional" avatars (e.g., anthropomorphized or a humanoid metaphor) is still to be investigated. Exploring this difference might help explain the results shown in this study.

6 CONCLUSION

We investigated the Great Chain of Being metaphor as a tool to traverse an agent metaphor space that extends beyond the human metaphor, which the HCI community has widely studied. Results from our study suggest that it is possible to understand non-human metaphors in a conceptually structured fashion. By being aware of the anatomy of metaphorical agent representation, designers and researchers studying human-agent interaction can more deliberately apply metaphors to an agent. To the best of our knowledge, this is the first study presenting evidence that worker engagement, intrinsic motivation, cognitive task load, and worker trust can be affected when conversational agents embody metaphors from different Great Chain of Being hierarchical categories. Our findings suggest that there is a potential trade-off in using metaphors from each level of being, e.g., some improve worker engagement while others decrease the cognitive task load. Specifically, using the lowest chain of being (an inorganic object metaphor — book) may significantly reduce the perceived workload of workers but may reduce intrinsic interest and enjoyment of the workers. In contrast to what exisiting HCI studies suggest, the human metaphor did not enhance user engagement or user trust towards the agent within the conversational crowdsourcing context. We therefore argue that the choice of the category from the Great Chain of Being with which to represent a given conversational agent, should be a deliberate design choice for crowdsourcing task designers.

ACKNOWLEDGMENTS

We would like to thank all the anonymous crowd workers for their participation in our study, and the reviewers for their valuable feedback. This work was partially supported by the TU Delft Design@Scale AI lab within the TU Delft AI Initiative, and in part by the 4TU.CEE UNCAGE project.

REFERENCES

[1] Jennifer L. Aaker, Emily N. Garbinsky, and Kathleen D. Vohs. 2012. Cultivating admiration in brands: Warmth, competence, and landing in the "golden quadrant". *Journal of Consumer Psychology* 22, 2 (2012), 191–194. https://doi.org/10.1016/j.jcps.2011.11.012

[2] Tahir Abbas, Ujwal Gadiraju, Vassilis-Javed Khan, and Panos Markopoulos. 2021. Making Time Fly: Using Fillers to Improve Perceived Latency in Crowd-Powered Conversational Systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. 2–14.

[3] Tahir Abbas, Vassilis-Javed Khan, Ujwal Gadiraju, Emilia Barakova, and Panos Markopoulos. 2020. Crowd of oz: a crowd-powered social robotics system for stress management. *Sensors* 20, 2 (2020), 569.

[4] Tahir Abbas, Vassilis-Javed Khan, Ujwal Gadiraju, and Panos Markopoulos. 2020. Trainbot: A Conversational Interface to Train Crowd Workers for Delivering On-Demand Therapy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 3–12.

[5] Alan Aipe and Ujwal Gadiraju. 2018. Similarhits: Revealing the role of task similarity in microtask crowdsourcing. In *Proceedings of the 29th on Hypertext and Social Media*. 115–122.

[6] Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. 2018. Searchbots: User engagement with chatbots during collaborative search. In *Proceedings of the 2018 conference on human information interaction & retrieval*. 52–61.

[7] Nancy Baym, Limor Shifman, Christopher Persaud, and Kelly Wagman. 2019. INTELLIGENT FAILURES: CLIPPY MEMES AND THE LIMITS OF DIGITAL ASSISTANTS. *AoIR Selected Papers of Internet Research* 2019 (2019).

[8] Aronté Marie Bennett and Ronald Paul Hill. 2012. The universality of warmth and competence: A response to brands as intentional agents. *Journal of Consumer Psychology* 22, 2 (2012), 199–204.

[9] Timothy W Bickmore, Laura M Pfeifer, and Brian W Jack. 2009. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1265–1274.

[10] Alan F Blackwell. 2006. The reification of metaphor as a design tool. *ACM Transactions on Computer-Human Interaction (TOCHI)* 13, 4 (2006), 490–530.

[11] Mark Blythe and Elizabeth Buie. 2014. Chatbots of the gods: imaginary abstracts for techno-spirituality research. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*. 227–236.

[12] Johanna Böttcher, Tabea Lüttmer, Ulrich Wagner, Frank Asbrock, and Maarten H. W. van Zalk. 2020. *Examining the structural validity of stereotype content measures - A preregistered re-analysis of published data and discussion of psosisble future directions*. Vol. 1. 0–2 pages.

[13] Luka Bradeško, Michael Witbrock, Janez Starc, Zala Herga, Marko Grobelnik, and Dunja Mladenić. 2017. Curious Cat–Mobile, Context-Aware Conversational Crowdsourcing Knowledge Acquisition. *ACM Transactions on Information Systems (TOIS)* 35, 4 (2017), 1–46.

[14] Cynthia Breazeal. 2003. Toward sociable robots. *Robotics and autonomous systems* 42, 3-4 (2003), 167–175.

[15] Elizabeth Buie. 2016. Transcendhance: a game to facilitate techno-spiritual design. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 1367–1374.

[16] John T Bush and Rachel M Balven. 2021. Catering to the crowd: An HRM perspective on crowd worker engagement. *Human Resource Management Review* 31, 1 (2021), 100670.

[17] Justine Cassell and Timothy Bickmore. 2000. External manifestations of trustworthiness in the interface. *Commun. ACM* 43, 12 (2000), 50–56.

[18] Wei Lun Cheong, Younbo Jung, and Yin-Leng Theng. 2011. Avatar: A virtual face for the elderly. In *Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry*. 491–498.

[19] Peng Dai, Jeffrey M Rzeszotarski, Praveen Paritosh, and Ed H Chi. 2015. And now for something completely different: Improving crowdsourcing workflows with micro-diversions. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 628–638.

[20] Ewart J De Visser, Samuel S Monfort, Ryan McKendrick, Melissa AB Smith, Patrick E McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied* 22, 3 (2016), 331.

[21] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, and Philippe Cudré-Mauroux. 2014. Scaling-up the crowd: Micro-task pricing schemes for worker retention and latency improvement. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

[22] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G Ipeirotis, and Philippe Cudré-Mauroux. 2015. The dynamics of micro-task crowdsourcing: The case of amazon mturk. In *Proceedings of the 24th international conference on world wide web*. 238–247.

[23] Graham Dove. 2020. Monsters , Metaphors , and Machine Learning. (2020), 1–17.

[24] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I" like" it, then I hide it: Folk Theories of Social Feeds. In *Proceedings of the 2016 cHI conference on human factors in computing systems*. 2371–2382.

[25] Shaoyang Fan, Ujwal Gadiraju, Alessandro Checco, and Gianluca Demartini. 2020. CrowdCO-OP: Sharing Risks and Rewards in Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–24.

[26] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[27] Susan T. Fiske, Amy J.C. Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology* 82, 6 (2002), 878–902. https://doi.org/10.1037/0022-3514.82.6.878

[28] Megan French and Jeff Hancock. 2017. What's the Folk Theory? Reasoning About Cyber-Social Systems. *SSRN Electronic Journal* (2017), 1–37. https://doi.org/10.2139/ssrn.2910571

[29] Ujwal Gadiraju and Stefan Dietze. 2017. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. 105–114.

[30] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM conference on Hypertext and social media*. 218–223.

[31] Bashar S Gammoh, Fernando R Jiménez, and Rand Wergin. 2018. Consumer attitudes toward human-like avatars in advertisements: The effect of category knowledge and imagery. *International Journal of Electronic Commerce* 22, 3 (2018), 325–348.

[32] Martin J Gannon. 2002. Cultural metaphors: Their use in management practice and as a method for understanding cultures. *Online Readings in Psychology and Culture* (2002).

[33] Martin J Gannon and Rajnandini Pillai. 2015. *Understanding global cultures: Metaphorical journeys through 34 nations, clusters of nations, continents, and diversity.* Sage Publications.

[34] Yang Gao, Zhengyu Pan, Honghao Wang, and Guanling Chen. 2018. Alexa, my love: Analyzing reviews of amazon echo. In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 372–380.

[35] Maia Garau, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, and M Angela Sasse. 2003. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 529–536.

[36] Susan A Gelman and Cristine H Legare. 2011. Concepts and folk theories. *Annual review of anthropology* 40 (2011), 379–398.

[37] Zohar Gilad, Ofra Amir, and Liat Levontin. 2021. The Effects of Warmth and Competence Perceptions on Users' Choice of an AI System. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[38] Marija Grishin, Jessica Yexin Li, Jenny G Olson, Surendra N Singh, Marija Grishin, Jessica Yexin Li, Jenny G Olson, Vladas Griskevicius, and Patti Williams. 2017. Choosing Unhealthy to Appear Warm : How Consumers Signal Personality Traits via Food Choice. 45 (2017), 634–635.

[39] Jonathan Grudin and Richard Jacques. 2019. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[40] Agneta Gulz. 2006. Visual Design of Virtual Pedagogical Agents : Naturalism versus Stylization in Static Appearance. *6th International Conference on Intelligent Virtual Agents, IVA 2006* 2006 (2006), 1–9.

[41] Jingya Guo, Jiajing Guo, Changyuan Yang, Yanjing Wu, and Lingyun Sun. 2021. Shing: A Conversational Agent to Alert Customers of Suspected Online-payment Fraud with Empathetical Communication Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[42] Magnus Haake and Agneta Gulz. 2009. A look at the roles of look & roles in embodied pedagogical agents - A user preference perspectivenet. *International Journal of Artificial Intelligence in Education* 19, 1 (2009), 39–71.

[43] Georgios Halkias and Adamantios Diamantopoulos. 2020. Universal dimensions of individuals' perception: Revisiting the operationalization of warmth and competence with a mixed-method approach. *International Journal of Research in Marketing* 37, 4 (2020), 714–736. https://doi.org/10.1016/j.ijresmar.2020.02.004

[44] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. All those wasted hours: On task abandonment in crowdsourcing. In *Proceedings of the twelfth ACM international conference on web search and data mining*. 321–329.

[45] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. The impact of task abandonment in crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering* (2019).

[46] Xu Han, Michelle Zhou, Matthew J Turner, and Tom Yeh. 2021. Designing Effective Interview Chatbots: Automatic Chatbot Profiling and Design Suggestion Generation for Chatbot Debugging. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[47] Sandra G. Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society* (2006), 904–908. https://doi.org/10.1177/154193120605000909

[48] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX. *Human Mental Workload. Advances in Psychology* 52 (1988), 139–183.

[49] Ting-Hao Kenneth Huang, Joseph Chee Chang, and Jeffrey P Bigham. 2018. Evorus: A Crowd-powered Conversational Assistant Built to Automate Itself Over Time. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 295.

[50] Ting-Hao Kenneth Huang, Walter S Lasecki, and Jeffrey P Bigham. 2015. Guardian: A crowd-powered spoken dialog system for web apis. In *Third AAAI conference on human computation and crowdsourcing*.

[51] Angel Hsing-Chi Hwang and Andrea Stevenson Won. 2021. IdeaBot: Investigating Social Facilitation in Human-Machine Team Creativity. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[52] Ray Jackendoff, David Aaron, George Lakoff, and Mark Turner. 1991. More Than Cool Reason: A Field Guide to Poetic Metaphor. , 320 pages. https://doi.org/10.2307/415109

[53] Mohammad Hossein Jarrahi, Will Sutherland, Sarah Beth Nelson, and Steve Sawyer. 2020. Platformic management, boundary resources for gig work, and worker autonomy. *Computer supported cooperative work (CSCW)* 29, 1 (2020), 153–189.

[54] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics* 4, 1 (2000), 53–71.

[55] Kari Daniel Karjalainen, Anna Elisabeth Sofia Romell, Photchara Ratsamee, Asim Evren Yantac, Morten Fjeld, and Mohammad Obaid. 2017. Social drone companion for the home environment: A user-centric exploration. In *Proceedings of the 5th International Conference on Human Agent Interaction*. 89–96.

[56] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020). https://doi.org/10.1145/3415234 arXiv:2008.02311

[57] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1301–1318.

[58] Tomoko Koda and Pattie Maes. 1996. Agents with faces: The effect of personification. In *Proceedings 5th IEEE International Workshop on Robot and Human Communication. RO-MAN'96 TSUKUBA*. IEEE, 189–194.

[59] Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*. Springer, 13–30.

[60] Gerald P Krueger. 1989. Sustained work, fatigue, sleep loss and performance: A review of the issues. *Work & Stress* 3, 2 (1989), 129–141.

[61] Tomasz P Krzeszowski. 1997. *Angels and devils in hell: Elements of axiology in semantics*. Energeia.

[62] Anastasia Kuzminykh, Jenny Sun, Nivetha Govindaraju, Jeff Avery, and Edward Lank. 2020. Genie in the bottle: Anthropomorphized perceptions of conversational agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[63] George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

[64] George Lakoff and Mark Turner. 2009. *More than cool reason: A field guide to poetic metaphor*. University of Chicago press.

[65] Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2010. Receptionist or information kiosk: how do people talk with a robot?. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 31–40.

[66] Eow Yee Leng, Roselan Baki, Rosnaini Mahmud, et al. 2010. Stability of the Intrinsic Motivation Inventory (IMI) for the use of Malaysian form one students in ICT literacy class. *Eurasia Journal of Mathematics, Science and Technology Education* 6, 3 (2010), 215–226.

[67] Ewa Luger and Abigail Sellen. 2016. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5286–5297.

[68] Richard Mander, Gitta Salomon, and Yin Yin Wong. 1992. A "pile" metaphor for supporting casual organization of information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 627–634.

[69] Panagiotis Mavridis, Owen Huang, Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2019. Chatterbox: Conversational interfaces for microtask crowdsourcing. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. 243–251.

[70] Edward McAuley, Terry Duncan, and Vance V Tammen. 1989. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport* 60, 1 (1989), 48–58.

[71] Edward D. McAuley, Terry Duncan, and Vance V. Tammen. 1989. Psychometric properties of the intrinsic motivation inventoiy in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport* 60, 1 (1989), 48–58. https://doi.org/10.1080/02701367.1989.10607413

[72] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a HIT: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2271–2282.

[73] Michael F McTear. 2016. The rise of the conversational interface: A new kid on the block?. In *International workshop on future and emerging trends in language technology*. Springer, 38–49.

[74] Robert J Moore, Raphael Arar, Guang-Jie Ren, and Margaret H Szymanski. 2017. Conversational UX design. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 492–497.

[75] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100.

[76] Ian Mull, Jamie Wyss, Eunjung Moon, and Seung-Eun Lee. 2015. An exploratory study of using 3D avatars as online salespeople: The effect of avatar type on credibility, homophily, attractiveness and intention to interact. *Journal of Fashion Marketing and Management* (2015).

[77] Donald A Norman. 1986. Cognitive engineering. *User centered system design* 31 (1986), 61.

[78] Kristine L Nowak. 2004. The influence of anthropomorphism and agency on social judgment in virtual environments. *Journal of Computer-Mediated Communication* 9, 2 (2004), JCMC925.

[79] Kristine L Nowak and Frank Biocca. 2003. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments* 12, 5 (2003), 481–494.

[80] Kristine L Nowak and Christian Rauh. 2005. The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction. *Journal of Computer-Mediated Communication* 11, 1 (2005), 153–178.

[81] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human Computer Studies* 112, December 2017 (2018), 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004

[82] Jonas Oppenlaender. 2020. Socially Augmented Crowdsourced Collection of Folk Theories. (2020).

[83] Amanda Purington, Jessie G Taft, Shruti Sannon, Natalya N Bazarova, and Samuel Hardman Taylor. 2017. " Alexa is my new BFF" Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 2853–2859.

[84] Sihang Qiu, Alessandro Bozzon, Max V Birk, and Ujwal Gadiraju. 2021. Using Worker Avatars to Improve Microtask Crowdsourcing. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 5. ACM New York, NY, USA, 1–28.

[85] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Estimating conversational styles in conversational microtask crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

[86] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[87] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Just the Right Mood for HIT!. In *International Conference on Web Engineering*. Springer, 381–396.

[88] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. TickTalkTurk: Conversational crowdsourcing made easy. *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* (2020), 53–57. https://doi.org/10.1145/3406865.3418572

[89] Jeffrey M Rzeszotarski, Ed Chi, Praveen Paritosh, and Peng Dai. 2013. Inserting micro-breaks into crowdsourcing workflows. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[90] Edward Schneider, Yifan Wang, and Shanshan Yang. 2007. Exploring the Uncanny Valley with Japanese Video Game Characters.. In *DiGRA Conference*.

[91] Donald A Schön. 1979. Generative metaphor: A perspective on problem-setting in social policy. *Metaphor and thought* 254 (1979), 283.

[92] Robin Sease. 2008. Metaphor's role in the information behavior of humans interacting with computers. *Information technology and libraries* 27, 4 (2008), 9–16.

[93] Verónica Sevillano and Susan T. Fiske. 2016. Warmth and competence in animals. *Journal of Applied Social Psychology* 46, 5 (2016), 276–293. https://doi.org/10.1111/jasp.12361

[94] Ameneh Shamekhi, Q Vera Liao, Dakuo Wang, Rachel KE Bellamy, and Thomas Erickson. 2018. Face Value? Exploring the effects of embodiment for a group facilitation agent. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[95] Muzafer Sherif, Daniel Taub, and Carl I Hovland. 1958. Assimilation and contrast effects of anchoring stimuli on judgments. *Journal of experimental psychology* 55, 2 (1958), 150.

[96] Ignacio Siles, Andrés Segura-Castillo, Ricardo Solís, and Mónica Sancho. 2020. Folk theories of algorithmic recommendations on Spotify: Enacting data assemblages in the global South. *Big Data & Society* 7, 1 (2020), 2053951720923377.

[97] Su Mae Tan, Tze Wei Liew, Chin Lay Gan, and Wee Ming Wong. 2021. Visual style of embodied virtual sales agents. *International Journal of Technology and Human Interaction* 17, 1 (2021), 1–13. https://doi.org/10.4018/IJTHI.2021010101

[98] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 77–87.

[99] Astrid M Von der Pütten, Nicole C Krämer, Jonathan Gratch, and Sin-Hwa Kang. 2010. "It doesn't matter what you are!" explaining social effects of agents and avatars. *Computers in Human Behavior* (2010).

[100] Belongie S. Wah C., Branson S., Welinder P., Perona P. [n.d.]. The Caltech-UCSD Birds-200-2011 Dataset. *Computation & Neural Systems Technical Report, CNS-TR-2011-001* ([n. d.]). https://doi.org/10.3182/20090902-3-US-2007.0059

[101] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[102] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. " Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 7–9.

[103] Alan Wexelblat. 1998. Don't make that face: A report on anthropomorphizing an interface. *Intelligent Environments* 173, 179 (1998), 98–02.

[104] Yorick Wilks. 2010. Is a Companion a distinctive kind of relationship with a machine?. In *Proceedings of the 2010 Workshop on Companionable Dialogue Systems*. 13–18.

[105] Aiyana K Willard and Rita A McNamara. 2019. The minds of god (s) and humans: Differences in mind perception in Fiji and North America. *Cognitive science* 43, 1 (2019), e12703.

[106] Luyan Xu, Xuan Zhou, and Ujwal Gadiraju. 2019. Revealing the role of user moods in struggling search tasks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1249–1252.

[107] Jie Yang, Judith Redi, Gianluca Demartini, and Alessandro Bozzon. 2016. Modeling task complexity in crowdsourcing. In *Fourth AAAI Conference on human computation and crowdsourcing*.

A WORKER ENGAGEMENT RESULT

Table 10. The Kruskal-Wallis test result across overall UES-SF score and subscales of UES-SF. Significant results are marked in italic bold. (α = 0.05, N=341)

	Overall UES	**Focused Attention**	**Perceived Usability**	**Aesthetic Appeal**	**Reward Factor**
Kruskal-Wallis H	11.20	10.05	13.57	12.12	7.94
df	5	5	5	5	5
sig	***.048***	.074	***.019***	***.033***	.159

Table 11. Overall UES-SF score Mann-Whitney post-hoc results between conditions. Significant results are marked in italic bold (Bonferroni-corrected alpha level 0.008). We calculated the effect size only when the statistical difference was significant. (*= all five metaphors)

Measure	**Metaphor**	**Compared metaphor group**	U	z	p	**effect size**
Overall UES-SF score	**God** (*Md* = 3.83, *n* = 55)	**Human** (*Md* = 3.83, *n* = 57)	1535.00	-.190	.850	-
		Animal (*Md* = 3.83, *n* = 56)	1277.00	-1.553	.120	-
		Plant (*Md* = 3.83, *n* = 59)	1584.00	-.219	.827	-
		Object (*Md* = 3.58, *n* = 57)	1374.00	-1.411	.158	-
		Control (*Md* = 3.83, *n* = 55)	1418.00	-.566	.572	-
	Human	**Animal**	1315.00	-1.617	.106	-
		Plant	1681.00	-.003	.998	-
		Object	1347.00	-1.851	.064	-
		Control	1504.50	-.367	.713	-
	Animal	**Plant**	1386.00	-1.491	.136	-
		Object	1046.00	-3.396	***.001***	moderate (r = .32)
		Control	1339.50	-1.185	.236	-
	Plant	**Object**	1425.50	-1.698	0.90	-
		Control	1566.50	-.318	.751	-
	Object	**Control**	1264.50	-2.033	0.42	-
	Control	**Agent Avatar*** (*Md* = 3.75, *n* = 286)	7494.00	-.555	.579	-

Table 12. Perceived Usability Mann-Whitney post-hoc results between conditions. Significant results are marked in italic bold (Bonferroni-corrected alpha level 0.008). We calculated the effect size only when the statistical difference was significant. (*= all five metaphors)

Measure	Metaphor	Compared metaphor group	*U*	*z*	*p*	effect size
Perceived Usability subscale	**God** (*Md* = 4.33, *n* = 55)	**Human** (*Md* = 4.67, *n* = 57)	1362.00	-1.241	.215	-
		Animal (*Md* = 4.67, *n* = 56)	1172.00	-2.258	.024	-
		Plant (*Md* = 4..67, *n* = 59)	1430.00	-1.127	.260	-
		Object (*Md* = 5.00, *n* = 59)	1282.00	-2.029	.043	-
		Control (*Md* = 5.00, *n* = 55)	1014.00	-3.193	***.001***	moderate (*r* = .30)
	Human	**Animal**	1457.00	-.838	.402	-
		Plant	1630.00	-.296	.768	-
		Object	1544.00	-.804	.422	-
		Control	1224.00	-2.162	.031	-
	Animal	**Plant**	1447.00	-1.197	.231	-
		Object	1651.50	-.003	.998	-
		Control	1300.00	-1.538	.124	-
	Plant	**Object**	1546.50	-1.098	.272	-
		Control	1201.50	-2.556	.011	-
	Object	**Control**	1401.00	-1.389	.165	-
	Control	**Agent Avatar*** (*Md* = 4.66, *n* = 286)	6140.50	-2.717	***.007***	lower than small (*r* = .06)

Table 13. Aesthetic Appeal Mann-Whitney post-hoc results between conditions. Significant results are marked in italic bold (Bonferroni-corrected alpha level 0.008). We calculated the effect size only when the statistical difference was significant. (*= all five metaphors)

Measure	Metaphor	Compared metaphor group	*U*	*z*	*p*	effect size
Aesthetic Appeal subscale	**God** (*Md* = 3.67, *n* = 55)	**Human** (*Md* = 3.33, *n* = 57)	1323.50	-1.429	.153	-
		Animal (*Md* = 3.67, *n* = 56)	1526.50	-0.80	.936	-
		Plant (*Md* = 3.33, *n* = 59)	1455.00	-.955	.340	-
		Object (*Md* = 3.33, *n* = 59)	1215.00	-2.324	0.20	-
		Control (*Md* = 3.33, *n* = 55)	1242.00	-1.626	.104	-
	Human	**Animal**	1252.50	-1.985	0.47	-
		Plant	1556.50	-.695	.487	-
		Object	1483.50	-1.101	.271	-
		Control	1484.50	-.486	.627	-
	Animal	**Plant**	1426.50	-1.269	.204	-
		Object	1115.00	-3.023	***.003***	small (*r* = .28)
		Control	1189.50	-2.078	0.38	-
	Plant	**Object**	1411.00	-1.784	.074	-
		Control	1383.00	-1.372	.170	-
	Object	**Control**	1526.00	-.550	.582	-
	Control	**Agent Avatar*** (*Md* = 3.33, *n* = 286)	7086.00	-1.170	.242	-

B INTRINSIC MOTIVATION: INTEREST-ENJOYMENT RESULT

Table 14. The Kruskal-Wallis test result across overall IMI score and subscales of IMI. Significant results are marked in italic bold. (α = 0.05, N=341)

	Overall IMI	**Interest-Enjoyment**	**Competence**
Kruskal-Wallis H	9.64	17.46	4.52
df	5	5	5
sig	0.86	***.004***	.478

Table 15. The post-hoc Mann-Whitney test of Interest-Enjoyment subscale from the IMI score. Significant results are marked in italic bold (Bonferroni-corrected alpha level 0.008). We calculated the effect size only when the statistical difference was significant. (*=all five metaphors together)

Measure	**Metaphor**	**Compared metaphor group**	U	z	p	**effect size**
Interest-Enjoyment	**God** (Md = 3.80, n = 55)	**Human** (Md = 3.80, n = 57)	1490.50	-.449	.653	-
		Animal (Md = 3.80, n = 56)	1472.00	-.403	.687	-
		Plant (Md = 3.60, n = 59)	1516.00	-.606	.545	-
		Object (Md = 3.20, n = 59)	1178.50	-2.524	.012	-
		Control (Md = 4.00, n = 55)	1428.50	-.504	.614	-
	Human	**Animal**	1415.00	-1.045	.296	
		Plant	1642.50	-.216	.829	-
		Object	1206.00	-2.634	***.008***	small (r = .24)
		Control	1353.00	-1.253	.210	-
	Animal	**Plant**	1416.00	-1.325	.185	-
		Object	1034.00	-3.470	***.001***	moderate (r = .32)
		Control	1516.00	-.142	.887	-
	Plant	**Object**	1308.50	-2.332	.020	-
		Control	1376.00	-1.403	.161	-
	Object	**Control**	974.00	-3.690	***.000***	moderate (r = .35)
	Control	**Agent Avatar*** (Md =, n = 286)	6647.50	-1.823	.068	-

C COGNITIVE TASK LOAD RESULT

Table 16. Kruskal-Wallis test result for the NASA-TLX. Significant results are marked in italic bold. (α = 0.05, N=341)

	TLX
Kruskal-Wallis H	23.74
df	5
sig	***.000***

Table 17. The post-hoc Mann-Whitney test of NASA-TLX. Significant results are marked in italic bold (Bonferroni-corrected alpha level 0.008). (*=all five metaphors together). We calculated the effect size only when the statistical difference was significant.

Measure	Metaphor	Compared metaphor group	U	z	p	effect size
NASA-TLX	**God** (*Md* = 29.17, *n* = 55)	**Human** (*Md* = 20.83, *n* = 57)	1271.00	-1.732	.083	-
		Animal (*Md* = 20.83, *n* = 56)	1157.50	-2.263	.024	-
		Plant (*Md* = 25.00, *n* = 59)	1310.00	-1.777	.076	-
		Object (*Md* = 12.50, *n* = 59)	925.00	-3.967	***.000***	moderate (*r* = .37)
		Control (*Md* = 16.67, *n* = 55)	937.00	-3.455	***.001***	moderate (*r* = .33)
	Human	**Animal**	1485.50	-.638	.524	-
		Plant	1668.00	-.075	.940	-
		Object	1160.00	-2.893	***.004***	small (*r* = .27)
		Control	1241.50	-1.909	.056	-
	Animal	**Plant**	1576.00	-.427	.669	-
		Object	1231.00	-2.365	.018	-
		Control	1326.00	-1.269	.204	-
	Plant	**Object**	1230.50	-2.755	***.006***	small (*r* = .25)
		Control	1307.00	-1.789	.072	-
	Object	**Control**	1393.50	-1.306	.192	-
	Control	**Agent Avatar*** (*Md* = 20.83, *n* = 286)	6663.00	-1.801	.072	-

D TRUST RESULT

Table 18. Kruskal-Wallis test result for the TiA. There was no significant difference between different metaphors ($\alpha = 0.05$, N=341).

	Trust	Propensity to Trust
Kruskal-Wallis H	2.482	3.013
df	5	5
sig	.779	.698

E OUTPUT ACCURACY AND TASK EXECUTION TIME RESULT

Table 19. Kruskal-Wallis test for the output accuracy and task execution time. There was no significant difference in output accuracy and task execution time between different metaphors.

	Accuracy	Execution time
Kruskal-Wallis H	6.386	8.299
df	5	5
sig	.270	.141