Aalborg Universitet

The Effect of Embodied Anthropomorphism of Personal Assistants on User Perceptions

Schneiders, Eike; Papachristos, Eleftherios; van Berkel, Niels

The Effect of Embodied Anthropomorphism of Personal Assistants on User Perceptions

Eike Schneiders
Department of Computer Science,
Aalborg University
Aalborg, Denmark
eike@cs.aau.dk

Eleftherios Papachristos
Department of Computer Science,
Aalborg University
Aalborg, Denmark
papachristos@cs.aau.dk

Niels van Berkel
Department of Computer Science,
Aalborg University
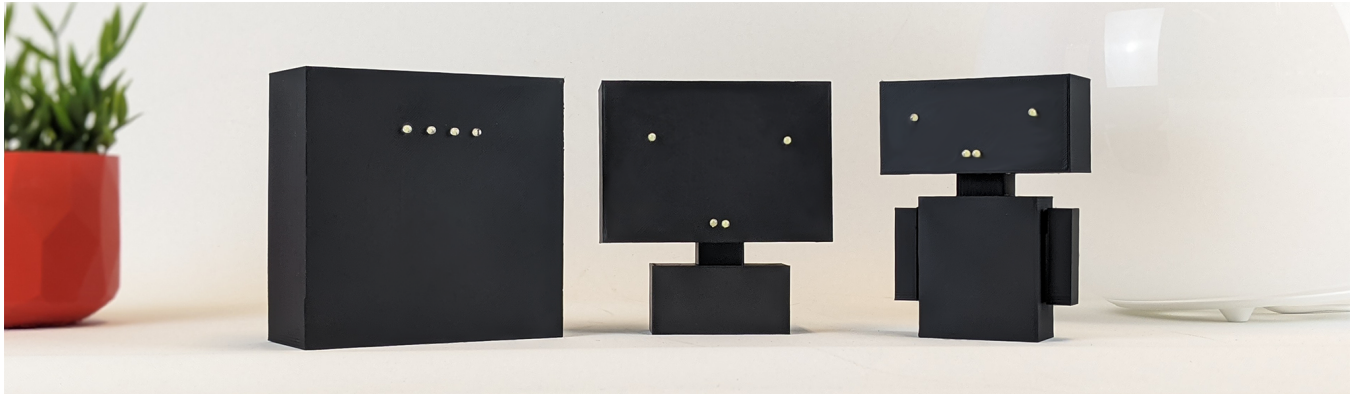Aalborg, Denmark
nielsvanberkel@cs.aau.dk

Figure 1: Overview of the three personal assistants presented in our study.

ABSTRACT

We investigate the impact of anthropomorphism on embodied AI through a study of personal assistants (PA). The effects of physical embodiment remain underexplored while the consumer market for PAs shows an increase in the diversity of physical appearances of these products. We designed three fictional personal assistants with varying levels of embodied anthropomorphism. We validated that our prototypes differed significantly in levels of anthropomorphism (N = 26). We developed a set of identical videos for each device, demonstrating realistic end-user interaction across six scenarios. Using a between-subject video survey study (N = 150), we evaluate the impact of different levels of embodied anthropomorphism on the perception of personal assistants. Our results show that while anthropomorphism did not significantly affect the perception of *Overall Goodness*, it affected perceptions of *Perceived Intelligence*, *Likeability*, and the device's *Pragmatic Qualities*. Finally, we discuss the implications of the identified relationships between anthropomorphism and user confidence in embodied AI systems.

CCS CONCEPTS

• **Human-centered computing → Interaction devices**; **Empirical studies in HCI**.

KEYWORDS

Digital assistants, Anthropomorphism, Likeability, Perceived intelligence, Physical embodiment

1 INTRODUCTION

Recent developments in Machine learning (ML) and Artificial intelligence (AI) have led to the widespread development and adoption of commercialised intelligent support systems. AI-driven applications are increasingly finding their way into a multitude of domains and contexts, including, for example, the home [13, 37], the work environment [8, 17], and transportation [23, 27]. Consequently, the average person interacts daily with a multitude of AI/ML-based applications while browsing the internet or using their music and video streaming services. Compared to behind-the-scenes recommendation systems or ordinary spell checkers, Personal Assistants (PA) are extremely visible to their end users, and have captivated the public attention as highlighted by the 157 million smart speakers sold in the US in 2019 alone [35]. PAs have evolved rapidly in their capabilities and appearance, while initially they were only available

as applications on smartphones, they can currently be encountered in a variety of shapes and forms. Although the embodiment of most commercially available PAs resembles a column-like speaker layout (*e.g.*, Google Home or Amazon Echo), a number of alternative approaches to the design of AI's physical embodiment have emerged. Examples include the Lynx [6] and Vector [20] robots with Alexa integration, as well as Jibo [39]. A common trend among these novel design approaches is the use of anthropomorphic features such as human-like eyes or limbs. While the tendency of using anthropomorphic features in the physical embodiment of AI assistants is increasing, the impact of these anthropomorphic features on user perceptions of the device and its capabilities remains underexposed in the Human-Computer Interaction (HCI) literature [45].

In this paper, we investigate the impact of physical anthropomorphism on peoples' perception of embodied AI. Previous research focusing on anthropomorphism has investigated aspects such as voice of *virtual* assistants in augmented reality [15], the effect of robot movement on user perception [14], and how the degree of anthropomorphism impacts interaction with virtual, on-screen, agents [16]. In contrast to existing research, we focus exclusively on the effect of anthropomorphism on user perception towards PAs. We aim to add to the growing body of research on anthropomorphism of AI-infused interactive technology by focusing on a specific device category. Research findings across different interactive technologies have shown that user perceptions of the physical design of devices and products can be shaped by various factors such as prototypicality [41], category [34], and context of use [43]. Prototypicality, which is defined as "*the amount to which an object is representative of a class of objects*" [21], is an important factor of visual design that can influence user acceptance of a device. In user perceptions, it may be more typical for robots to have anthropomorphic features such as arms or eyes compared to PAs which, at least until recently, were primarily looking like speakers. Therefore, we consider it important to study the effects of anthropomorphism on user perceptions within the specific and well-defined product category of smart speakers (or embodied PAs) rather than a more generic and abstract class of AI-enabled devices.

In order to investigate anthropomorphism in the physical embodiment of PAs, we conduct two studies. First, we design and present three alternative representations of our Fictional Personal Assistant (FiPA) with an increasing degree of anthropomorphism. Second, we conduct a verification study (N = 26), to verify that we operationalised anthropomorphism in a satisfactory manner since participants ratings showed statistically significant differences among the three prototypes. Following this, we conduct an online survey study [12, 40] in which we showcase videos of our three personal assistants across six typical usage scenarios in a between-subject study (N = 150). In order to provide a realistic evaluation, our scenarios include three successful and three failed interactions. We investigate the impact of the personal assistant's physical embodiment on Perceived Intelligence, Likeability, Overall Goodness, as well as Attractiveness measured by Pragmatic, and Hedonic qualities.

Our results reveal a statistical significant impact of embodied anthropomorphism on the PAs Likeability, Perceived Intelligence, as well as Pragmatic qualities. The findings from our study inform the design of future personal assistants and present implications for future work on Human-AI interaction.

2 RELATED WORK

The HCI community has a long history of studying the effect of systems' external characteristics on user interaction. Already at CHI 1994, Walker et al. report on the effect of embedding a talking human face in a digital questionnaire application, with their results pointing to higher levels of end-user engagement as compared to a text-based display [46]. A more recent example can be found in the work by Knijnenburg & Willemsen, who study interaction with an agent providing computer-like cues and human-like appearance (and capability) cues [16]. Their results highlight how the agent's appearance can affect a user's mental model of the system's operation and expectations towards its abilities. Kuramoto et al. demonstrate that conversational agents used in customer support can be designed to suppress customer's anger by following a Balance Theory-informed conversational style [18].

Within the domain of virtual assistants (VA) and personal assistants (PA), prior work has investigated the impact of various anthropomorphic aspects [15, 36, 45, 48]. In a recent study, Kim et al. [15] investigate the impact of embodiment in augmented reality (AR) on intelligent virtual assistants. While most virtual assistant are only represented through the use of voice, Kim et al. conduct an experiment in which participants interacted with a virtual assistant in three distinct conditions, (1) a disembodied voice, (2) a virtual model of a human using upper body gestures, as well as (3) a virtual model of a human that uses locomotion in addition to speech and upper body gestures. They concluded that the virtual embodiment resulted in an increase in the users' confidence in the virtual agents' ability to manipulate real-life object such as turning on the lights. Furthermore, the presence of embodiment increased the users' sense of engagement, social richness, as well as social presence with the virtual assistant. These findings point towards an added benefit of a virtual embodiment as compared to a disembodied voice, when interacting with a virtual assistant through AR.

Focusing on the use of PAs in a cooking context, Sano et al. explore the use of onomatopoetic expressions to guide users [36]. Onomatopoetic expressions are phonetic utterances (*e.g.*, 'meow' in English for vocalisation of a cat), commonly used in Japanese language in cooking-related activities. Their results highlight how PAs do not necessarily need a strict or neutral expression to be useful and accepted by end-users. Finally, Joosse et al. studied the effect of the movement of embodied PAs on user perceptions [14]. Their results highlight that while robots may not be able to perfectly mimic human approach styles, they are able to mitigate these negative effects by more clearly communicating their intent (*e.g.*, decreasing motor noise indicates slowing down).

Yang and Lee investigate what factors impact potential customers intention to adopt and use virtual personal assistants such as the Apple HomePod, Amazon Echo, or the Google Home [48]. The authors investigate the impact of both 'provided functionality' and the 'physical embodiment' of the device itself and its importance to the users' choice to adopt and use the personal assistant. Through a survey study, Yang and Lee confirm the importance of functional

factors such as perceived usefulness or willingness to adopt and use, but also identify a significant impact of the physical embodiment, measured as visual attractiveness on participants adoption and usage intention.

As our discussion of prior work highlights, anthropomorphism can be implemented using multiple different approaches, including embodiment, movement, authenticity of voice, as well as the ability to have a natural, human-like dialogue. Wagner et al. demonstrate significant impact of anthropomorphic features on the perception of users on voice assistants [45]. Using an online survey they investigated several different aspects in voice assistants including 'perceived sociability', 'usage intention' as well as 'likeability'. The survey included only participants with prior experience of personal assistants. Wagner et al. identify a significant role of anthropomorphism for personal voice-based assistants, such as higher positive impressions of the PA through anthropomorphism, which again leads to an increased intention of use.

These prior studies highlight the effect of various anthropomorphic aspects on perceptions towards PAs and VAs. In this work, we investigate the effect of *embodied* anthropomorphism on PAs.

3 VALIDATION STUDY: EVALUATION OF ANTHROPOMORPHISM

Our goal in this study was to validate that the three FiPA prototypes represented distinct levels of anthropomorphic embodiment. To do so, we conducted an online survey study in which 26 participants were presented with still images of all three versions, thereby removing any influence interaction could have effects, as we were interested in differentiating the devices purely by their *physical embodiment*.

3.1 Method

3.1.1 Participants. We recruited a total of 26 participants (16 male, 10 female, age range 22–46, average age = 29.8, SD = 5.54). Participants were recruited using the authors' personal networks, as well as through social media. We did not compensate participants for their participation.

3.1.2 Devices. For study stimuli, we decided to design our own prototypes that would resemble the appearance of PAs. We discarded the idea of using existing commercially available PAs as this could introduce preferences towards specific devices based on brand recognition [1]. We also considered but discarded the idea of using other devices that resemble embodied AI (*e.g.*, security cameras, domestic robots, or toys) since these devices were developed for a specific purpose, this could introduce confounding factors such as visual design characteristics typical in one product category but not in the physical design of PAs. In order to, (1) minimise the influencing factors of brand and design characteristic from other product categories, (2) guarantee *comparability between devices* in terms of production quality, and to (3) ensure that participants had *no previous experience* with the personal assistants evaluated in the study, we 3D-modelled and printed three bespoke personal assistant with different degrees of anthropomorphism (see Figure 2). To reduce the effect of potential confounding factors as much as possible, we focused exclusively on a straightforward manipulation of the devices' embodiment (*i.e.*, the physical shape). Other

physical aspects, such as the material, height, number of LEDs, and colour were kept identical between devices. Our design goal was to uphold an as simple as possible appearance, thereby reducing the risk of introducing confounding factors and allowing us to obtain a gradual increase in anthropomorphic features. When designing, we started with the simplest version of FiPA (Prototype A), which design was inspired by existing solutions, resembling a column with four centrally placed LED lights for feedback. In order to increase the level of anthropomorphism for Prototype B, we sought inspiration in existing products[1] who rely on the simple use of facial expression (*i.e.*, LED placement to resemble eyes or mouth) to anthropomorphism the device. We used a similar approach for the design of Prototype C. Prior to designing the prototypes in Blender[2], we sketched detailed versions of the prototypes, including accurate dimensions.

3.1.3 Procedure. Each of the study participants was given a link to an online survey webpage. On the first page of the survey website, we provided information about the purpose of the study and asked participants to fill in a consent form. The study was anonymous, and we only asked participants to provide basic demographic information (i.e., age, gender, experience with PAs) that we considered possible factors that could influence the results. We then presented our participants with a definition of 'Anthropomorphism' as well as examples to make sure that there was a common understanding of the term. For easy reference, this definition was visible at the top of the survey page at all times and was phrased as follows:

> *Objects, shapes, or forms that appear to be human in appearance, character, or behaviour are considered anthropomorphic. In other words, objects that have been made to resemble human form or characteristics.*

Following this, participants were presented with each of the prototypes individually and were asked to rate them first on a nine-point Likert scale ranging from 'Not anthropomorphic' (1) to 'Very anthropomorphic' (9) and then on the five-item Godspeed I questionnaire on anthropomorphism [3]. Since the Godspeed I questionnaire was developed for the evaluation of robots and not PAs we modified the last question in from a 'Moving rigidly' - 'Moving elegantly', to a 'I believe this prototype communicates: Rigidly - Elegantly' scale, following earlier work by Laban and Araujo [19]. Finally, we asked our participants to provide us with some feedback in a free form text field asking "*What words would you use to describe this prototype?*".

3.2 Results

The purpose of this validation study was to determine whether our prototypes differed in terms of anthropomorphism. To analyse the results, we first performed a repeated-measures ANOVA with the prototype as a three-level independent variable and anthropomorphism (measured on a nine-point scale) as the dependent variable. Results showed that the differences in anthropomorphism among the prototypes were highly significant ($F(2,50) = 248.1$, $p < .01$, $\eta^2 = .91$). As expected, based on user ratings the most anthropomorphic

---

[1] *e.g.*, https://m.media-amazon.com/images/I/51QrLvHZ3vL._AC_SL1200_.jpg
[2] https://www.blender.org

Figure 2: The three versions of FiPA with increasing degree of anthropomorphism (left to right). All aspects, apart from the physical embodiment, such as height, material, lightning, and the framing were kept consistent. Prototype A was 11x11cm (*w* x *h*), Prototype B was 10x11cm, and Prototype C was 8x11cm.

prototype was Prototype C (M = 7.73, SD = 1.4), followed by Prototype B (M = 5.54, SD = 1.5), and Prototype A (M = 1.38, SD = .64), see Figure 2. Post-hoc comparison with Bonferroni corrections furthermore revealed that all pair differences here were significant at a $p < .01$ level. We also investigated whether the participant variables of gender, age, or previous experience with PAs influenced these results. A mixed model repeated measures ANOVA with demographic variables as between-subjects factors showed no significant interaction effect of gender or previous experience with PAs and anthropomorphic assessment of the prototypes. However, we found a significant interaction effect of age and anthropomorphic assessment ($F(2,48) = 3.67$, $p = .03$, $\eta^2 = .13$). A post-hoc comparison did not reveal any significant differences between pairs of prototypes.

The analysis of the Godspeed I questionnaire data provided similar results. Before calculating average anthropomorphism scores for each prototype, we examined the internal consistency of the questionnaire, which was relatively high (Cronbach's α = .847). We first calculated an average anthropomorphism score for each device for each participant and then performed a repeated-measures ANOVA analysis on the aggregated data. Results showed that the differences in anthropomorphism among the prototypes were highly significant ($F(1.401, 35.01) = 60.08$, $p < .01$, $\eta^2 = .71$). Greenhouse-Geisser corrections were applied to within-subject effects to compensate for violations of sphericity. Aggregated scores of the Godspeed I questionnaire revealed the same ranking order as the previous analysis showing Prototype C as the most anthropomorphic (M = 3.42, SD = .92), followed by Prototype B (M = 2.68, SD = .71), and Prototype A (M = 1.87, SD = .62). Post-hoc comparison with Bonferroni corrections showed that all pair differences here are significant at a $p < .01$ level. We also performed mixed model repeated measures ANOVA with gender, age, and experience with PAs as between-subject factors to assess the stability of those results across demographic characteristics. This analysis showed no significant interaction effect between anthropomorphic assessment of our prototypes and any of the demographic variables.

We also asked participants to describe the prototypes in their own words. Analysis of these data revealed a clear pattern that aligned with questionnaire results. Words used to describe Prototype A were similar to those describing a machine. Examples include adjectives such as "Boxy, Square", "Machine-like" to full descriptions such as "looks like a router or an old radio" or "It reminds me a lot of my WiFi router at home" which confirms the findings of very low anthropomorphism for prototype A. In contrast,

Prototype C was described using adjectives and characteristics typical used to describe humans such as "human-like", "approachable" and "aware" and statements such as "...reminds me of Danbo the cardboard figure..." or "Box with eyes and cute tiny little arms...".

Results of the data analysis confirm that we operationalised anthropomorphism in a satisfactory manner and that our three prototypes represent significant distinct levels of anthropomorphic features.

4 STUDY: EFFECT OF ANTHROPOMORPHISM ON PERCEPTIONS TOWARDS PAs

In this study we evaluate the effect of anthropomorphism on participants' perceptions of a personal assistant through an online survey-based study. We used the same FiPA prototypes as introduced in Section 3, as these had significant differences in anthropomorphism. In a between-subject study, participants were asked to watch six videos in which a user interacts with FiPA across typical personal assistant-based tasks. We investigate how participants perceptions of the three FiPA prototypes differed in regards to *Overall Goodness*, *Likeability*, *Perceived Intelligence*, as well as *Attractiveness*.

4.1 Method

4.1.1 Participants. We recruited a total of 202 participants using Amazon Mechanical Turk. This was followed by the exclusion of participants that (1) failed to answer a simple control question, (2) pasted spam content in the free-text field, or (3) completed the survey in less time than technically possible given the length of the presented videos. This left us with a total of 150 valid responses. Our final selection of participants had an average age of 36.1 (23–69, SD = 10.88), 107 are male, 42 female, and one preferred not to disclose their gender. Of the final sample, 42 (28%) participants reported owning a voice assistant. Participants received 1.5$ for participation. Median completion duration for the entire questionnaire was 7 minutes and 41 seconds.

4.1.2 Measurements. We investigated the dependent variables of Overall Goodness, Perceived Intelligence, Likeability, Hedonic, and Pragmatic Qualities to assess participant variability in perceptions of our three prototypes. Overall Goodness assessments were measured on visual analogue single-item scales after each video and on a seven-point scale at the end of the study. Likeability and Perceived Intelligence were measured using the five-items Godspeed III and Godspeed IV questionnaires [3]. While these questionnaires

Example of a successful interaction
(https://youtu.be/UHgKEOwsVPY)

USER: FiPA activate.
FIPA: *Entering listening mode.*
USER: Turn on the Living room light.
FIPA: *Processing mode.*
FIPA: Alright, the living room lamp is now turned on.
FIPA: *Entering listening mode.*
USER: Turn the lamp blue.
FIPA: *Processing mode.*
FIPA: *Changing lamp light to blue.*
FIPA: Got it! The lamp is now blue.
FIPA: *Entering listening mode.*
USER: Make it brighter.
FIPA: *Processing mode.*
FIPA: *Increasing lamp brightness.*
FIPA: Done, I have increased the living room brightness.

Example of a failed interaction
(https://youtu.be/U5htPNFT9u4)

USER: FiPA activate.
FIPA: *Entering listening mode.*
USER: Play my breakfast music playlist on youtube.
FIPA: *Processing mode.*
FIPA: Sure thing, playing a breakdance music playlist from youtube.
FIPA: *Entering listening mode.*
USER: No not breakdance, breakfast.
FIPA: *Processing mode.*
FIPA: Do you want to order breakfast?
FIPA: *Entering listening mode.*
USER: No, play *breakfast music*!
FIPA: *Processing mode.*
FIPA: The breakdance music is already playing, do you want to start it over again?
FIPA: *Entering listening mode.*
USER: Nevermind, just turn of the music.

Figure 3: Example of two interaction flows as illustrated in the videos. *Left:* a successful interaction. *Right:* a failed interaction.

have been developed to evaluate robots, they have been used in previous research studies for the evaluation of other AI-based systems, such as conversational agents and virtual avatars (*e.g.*, [2, 4]). Attractiveness was measured using the AttrakDiff questionnaire [9] which makes a distinction between Pragmatic and Hedonic Qualities. Pragmatic Qualities refer to perceptions of usefulness and usability, while Hedonic Qualities refer to non-utilitarian values such as aesthetics and pleasurable experiences. For this study, we used the short four-items per factor version of the AttrakDiff questionnaire [10]). It has been found that this version of AttrakDiff outperformed other similar short version questionnaires [33].

4.1.3   Video Design. In order to investigate the impact of the physical embodiment of our FiPA prototypes on participant perceptions, we conducted an online survey in which we presented videos showing typical PA interaction scenarios. We produced identical videos for each prototype to ensure consistency. The use of crowdsourcing in combination with an online survey using videos are common data collection approaches in HCI/HRI, (see *e.g.*, Jensen et al. [12] or Tennent et al. [40]). By using video instead of actual interaction with the device, we maintain the highest degree of control over the similarity of interaction between participants and devices, thereby increasing the comparability between prototypes. Using natural language to interact with PAs often leads to voice recognition errors that would not be consistent among participants and therefore leading to variability of user experiences and therefore evaluation [24].

We developed six videos for each FiPA version, each showcasing the same three successful and three failed interactions. For all 18 videos we paid particular focus on consistency in regards to the demonstrated functionality. Further, we made sure that both the human interlocutor as well as the voice of FiPA was the same for all three conditions, which was guaranteed by using the same audio clips for each device-interaction combination. Lastly, we made sure

that the background, the LED animations, and the lightning was consistent across all 18 videos.

All three devices make use of the same light pattern, implemented using four white LEDs and an Arduino Uno, to demonstrate listening, processing, and providing feedback. The 'listening mode' is visualised as a constant glow of all four LED's, followed by the 'processing' visualisation in which the four LED's consecutively blink with a 140ms off-set, creating a wave like pattern. The 'feedback mode' is visualised by a synchronous pulse of all four LED's. In order to ensure a clear contrast between FiPA and the user in the videos, we recorded a human female voice to represent the user and made use of the British male Siri voice to generate an stereotypical automated male voice.

4.1.4   Procedure. Prior to enrolling to the study, participants were presented with a short text describing the purpose of the study. After signing a consent form they were randomly assigned to one of the three conditions that corresponded exclusively to one of the prototypes. Each participant watched six videos with the assigned prototype (see Table 1 for an overview of the six presented videos), which included three successful[3] and three failed interactions[4]. Transcripts of one successful and one failed interaction with FiPA, are provided in Figure 3. For reasons of ecological validity we chose to include use-cases showing both successful and failed interactions with FiPA prototype. Videos were presented in random order. The chosen successful scenarios are based on the three most common tasks performed with personal assistants as identified by Paay et al. [32], namely (1) playing media, (2) interacting with smart home accessories, and (3) creating reminders/events (see Table 1). The errors that were demonstrated in the failed scenarios were inspired by the error types presented by Myers et al. [31]. Following each

---

[3]See https://youtu.be/UHgKEOwsVPY for an example of a successful interaction.
[4]See https://youtu.be/U5htPNFT9u4 for an example of a failed interaction.

ID	Task	Duration	Outcome
1	Turn on light	29 sec.	Success
2	Play music	37 sec.	Success
3	Create reminder	32 sec.	Success
4	Turn on light	19 sec.	Unfamiliar intent (cannot parse / unsupported)
5	Play music	37 sec.	NLP error (maps utterance to wrong intent)
6	Create calendar entry	34 sec.	Failed feedback (ambiguous verbal feedback)

Table 1: Presented video scenarios and respective length, as inspired by [31, 32].

video, participants were asked to rate the presented interaction on a slider ranging from 'Bad' to 'Good' (0–100).

Following the sixth video, we asked them to fill in a final questionnaire consisting of 24 questions. These questions were comprised of five items for Likeability and Perceived Intelligence, four items for Hedonic and Pragmatic Qualities, and one question about Overall Goodness. In addition, the questionnaire also included a control question to assess attentiveness and four questions regarding the participants' demographic characteristics. The demographic questions were, assessing participants' ownership of a PA, usage frequency, gender, and age. We also provided an optional free-text input field for additional feedback. The questionnaire was deployed through Qualtrics.

4.2 Results

Our data analysis aimed to examine whether anthropomorphic embodiment could influence user perceptions of a PA. First, we tested whether the random allocation of participants to the condition resulted in groups with significant differences regarding demographic representation. Results showed that there were no significant differences among the three groups in in terms of gender ($\chi^2(4) = 6.34$, $p = .175$), PA ownership ($\chi^2(2) = 1.59$, $p = .452$), or previous experience with PA's ($\chi^2(4) = .98$, $p = .613$). Likewise, a Kruskal-Wallis test showed no groups differences regarding age (H(2) = 5.49, $p = .06$). These results let us conclude that the three groups that were randomly formed were similar regarding demographic characteristics.

The independent variable in this between-subject study was '*Anthropomorphism*', which had three levels ('Low', 'Moderate', and 'High') corresponding to the three different PA prototypes we developed. In our validation study (see Section 3), we confirmed our hypothesis that those prototypes varied significantly in regard to anthropomorphism. Our dependent variables were 'Overall Goodness', 'Perceived Intelligence', 'Likeability', 'Hedonic', and 'Pragmatic qualities'. We calculated average scores and standard deviations for all dependent variables for each prototype independently. Average raw ratings can be seen in Table 2, while Figure 4 shows box plots with normalised values to allow for an easier comparison of the results. A simple examination of average scores revealed that prototype C outperformed the other prototypes on all factors, with the only exception being the Hedonic Qualities, in which prototype A performed slightly better than prototype C. In contrast, prototype B was consistently the worst performing on all factors. The next step of our analysis was to inspect our data for violations of normality before performing statistical tests to identify possible

significant differences among the three conditions in regard to our dependent variables.

We performed Shapiro–Wilk and Kolmogorov-Smirnov tests, which, as expected, showed that none of our dependent variables was following a normal distribution. Therefore, we used Kruskal-Wallis tests with Dunn-Bonferroni post-hoc tests to assess the prototype's main effect on the dependent variables. We used the Rank Transform method [47] for factorial designs to investigate the influence of the participant variables (*e.g.*, PA ownership, gender, or age) on our dependent variables. Using this method requires transforming the data into ranks and subsequently perform factorial ANOVAs for all dependent variables. The following subsections provide a description of our analysis and results for each dependent variables individually.

4.2.1 Overall Goodness. The Goodness of FiPA was evaluated seven times throughout this study. Participants provided a Goodness rating after viewing each of the six videos on a 100-point visual analogue scale with the labels bad-good as anchors. At the end of the study, each participant was prompted to evaluate the prototypes on 'Overall Goodness' one last time on a seven-point Likert scale. The purpose of the final evaluation was to measure participants overall perception based on all the interactions they observed. In addition, the intermediate goodness ratings were collected to allow us to evaluate the effect of failed versus successful scenarios and the impact of scenario types (*i.e.*, Light, Music, Reminder/Event) on goodness perceptions.

We performed a Kruskal-Wallis test with Overall Goodness as the dependent variable and Anthropomorphism as a three-level independent variable (corresponding to the three prototypes). Our analysis showed that even though the most anthropomorphic prototype scored higher than the others (M = 6.21, SD = 1.61), the results were not statistically significant. We also performed a similar Kruskal-Wallis test with aggregated intermediate goodness ratings, which also showed non-significant results. These findings indicate that Anthropomorphism does not have a significant effect on Overall Goodness evaluations of PAs.

The next step of our analysis was to examine whether Goodness perceptions were affected by scenario type. As mentioned in Section 4.1.4, each participant observed interactions with FiPA in three distinct scenarios (Light, Music, Reminder/Event). For each of those scenarios, we showed one successful and one failed interaction. Since these are within-subjects measurements, we first aggregated Goodness ratings across scenarios and prototype Combinations and then conducted a Friedman test. The test revealed that there was a statistically significant difference ($\chi^2(2) = 23.68$,

Dependent Variable	Prototype A	Prototype B	Prototype C	Significance	η^2
Overall Goodness	5.92 (1.76)	5.79 (1.65)	6.21 (1.61)	ns.	-
Perceived Intelligence	3.66 (0.83)	3.31 (0.82)	3.73 (0.81)	$p = .029$	.035
Likeability	3.91 (0.64)	3.65 (0.76)	3.99 (0.75)	$p = .038$	.031
Pragmatic Qualities	5.79 (1.62)	5.33 (1.52)	6.15 (1.52)	$p = .010$	.049
Hedonic Qualities	5.9 (1.37)	5.35 (1.68)	5.86 (1.71)	ns.	-

Table 2: Data for all three FiPA prototypes.

$p < .01$) in how participants rated Goodness of FiPA in the three scenarios. The scenario in which goodness ratings were highest was Reminder/Event (Mdn = 77.2), followed by Light (Mdn = 58.2), and Music (Mdn = 54.2). Post-hoc analysis with Wilcoxon signed-rank tests showed that these differences were statistically significant at the $p < .01$ level. These results show that participants goodness perceptions were influences by the context of use (scenario type) of FiPA.

Next we investigated whether there were differences in user perceptions of FiPA based on the observed scenario's success level. To be able to compare, we first aggregated goodness ratings over the three successful and then the three failed scenarios. To no surprise, participants rated Goodness higher in successful (Mdn = 88) compared to failed scenarios (Mdn = 62.5). A Mann–Whitney test indicated this difference to be statistically significant, U(450) = 45353, z = -14.4, $p < .001$.

The final step of our analysis concerning Goodness was to investigate the possible effect of Anthropomorphism on the severity of participant judgement in failed scenarios compared to successful ones. A simple comparison of averages shows that the most anthropomorphic prototype received the highest ratings in successful (M = 86, SD = 12.1) and the lowest ratings in the unsuccessful scenarios (M = 46.6, SD = 35). The reverse trend was observed for the least anthropomorphic prototype, where we found the lowest average ratings for successful scenarios (M = 83.8, SD = 13.4) and the highest for unsuccessful ones (M = 52.9, SD = 32.9). To examine whether our data support the hypothesis that Anthropomorphism can affect severity ratings, we calculated the difference between ratings in successful and failed scenarios for each participant. We performed a one way ANOVA with Anthropomorphism as the independent and rating variance as the dependent variable. Results showed that rating variance was not significantly different among the three prototypes. Hence, we did not find support for the hypothesis that participants would be less forgiving towards Anthropomorphic embodied PAs.

4.2.2 Perceived Intelligence. For the dependent variable of Perceived Intelligence, we gathered user responses on the five-item Godspeed VI questionnaire. Initial analysis of the results showed that the scale's internal consistency was high (Cronbach α = 0.86). We then conducted a Kruskal-Wallis test with Anthropomorphism as the independent and Perceived Intelligence as the dependent variable. The results show that our participants perceptions of Intelligence for PAs differed significantly (H(2) = 7.59, $p = .022$, η^2 = .035) among the three conditions. prototype C, the most anthropomorphic, was perceived to be the most intelligent (Mdn = 3.8), followed by prototype A (Mdn = 3.6) and prototype B (Mdn = 3.2). Bonferroni

adjusted pairwise comparisons showed that the difference between Prototypes B and C was statistically significant ($p = .022$). These results show an effect of Anthropomorphism on perceptions of intelligence, but the relationship does not appear to be linear.

4.2.3 Likeability. Similar to Perceived Intelligence, to measure the dependent variable of Likeability, we used a five-item factor from the Godspeed Questionnaire III [3]. Internal consistency of the scale was high for Likeability, too (α = 0.81). The Kruskal-Wallis test with Anthropomorphism as the dependent and Likeability as the independent variable showed significant differences in how participants perceived the prototypes on this dimension (H(2) = 6.87, $p = .032$). Adjusted pairwise comparisons also showed significant differences ($p = .028$) only between prototype B (Mdn = 3.8) and prototype C (Mdn = 4). Similar to Perceived Intelligence, our results show that Anthropomorphism influenced the Likeability of FiPA but not in a linear way.

4.2.4 Pragmatic and Hedonic Qualities. To assess whether there were differences in participant perceptions of prototype attractiveness, we gathered participant ratings on the AttrakDiff questionnaire [10]. The results show that internal consistency was high for both hedonic (α = 0.84) and pragmatic (α = 0.86) factors of this questionnaire. Subsequent Kruskal-Wallis test with Anthropomorphism as the dependent and Perceived Intelligence as the independent variable and Hedonic qualities as the dependent did not show statistically significant difference. However, we found a significant effect of Anthropomorphism on the Pragmatic Qualities of the prototypes (H(2) = 6.87, $p < .01$, η^2 = .049). Post hoc comparisons showed only significant differences (p < .01) between Prototypes B (Mdn = 5.5) and C (Mdn = 6.5) regarding Pragmatic Qualities.

4.2.5 Effect of Participant Variables. To assess whether PA ownership was an influencing factor in our results, we performed factorial ANOVAs with Prototype and PA Ownership as independent variables after rank transforming our dataset. Even though we did not find any interaction effects between Prototype and PA Ownership, we found significant main effects of Ownership on Overall Goodness (F(1,149) = 5.76, $p = .02$), Likeability (F(1,149) = 8.01, $p < .01$), Perceive Intelligence (F(1,149) = 4.40, $p = .04$), Hedonic (F(1,149) = 4.86, $p = .03$), and Pragmatic Quality (F(1,149) = 5.09, $p = .03$). For these factors, participants who had a PA gave on average higher ratings to FiPA than those that did not own one.

4.2.6 Summary of Statistical Findings. Our results show that anthropomorphic features did not influence the evaluation of the overall Goodness of PAs, but it affected their Likeability and, more
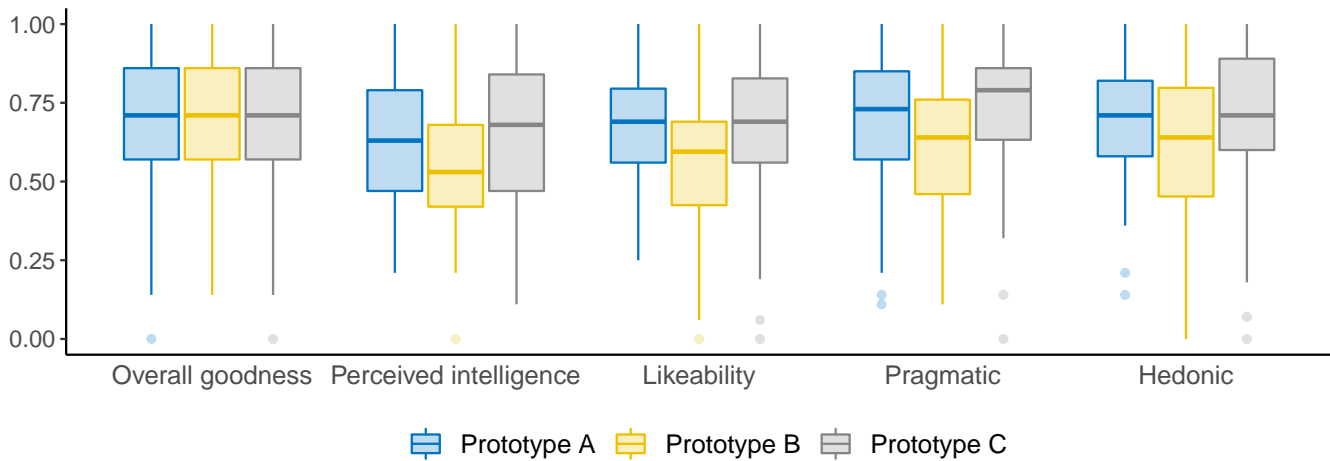
Figure 4: Overview of participant evaluation per prototype. The y-axis shows normalised values for comparison.

surprisingly, how intelligent they were perceived to be. Participants did not perceive the Hedonic but the Pragmatic qualities as significantly different among the prototypes with varying anthropomorphic features.

5 DISCUSSION

In this paper, we investigated the effect of anthropomorphic embodiment on users' perception of the PA. We designed three versions of the fictional personal assistant (FiPA), that only differed in physical embodiment. In our validation study, we confirmed that these three FiPA prototypes were significantly different in their level of anthropomorphism. Following this, we conducted an online video survey study with 150 participants. Participants were randomly assigned one of the three prototypes and presented with six videos showcasing three successful and three failed interactions with the respective version of FiPA. While our results show that anthropomorphism did not affect 'Overall Goodness', it did affect perceptions of 'Perceived Intelligence', 'Likeability', and the 'Pragmatic Qualities'.

In the following sections we provide a discussion of our findings focusing on implications for design, expectation management, as well as a potential uncanny valley [30] effect for personal assistants.

5.1 Designing for AI-Assistance

While the topic of PAs has received an increased focus in the HCI community, only a third of US adults own a smart speaker [44]. It is, therefore, important that HCI researchers carefully consider how AI-powered assistive devices can successfully enter user's living space. One aspect of this challenge is the physical appearance of PAs. Our results reveal that Perceived Intelligence, Likeability, and Pragmatic quality are all significantly affected by the embodiment of the PA – with the PA with the highest level of anthropomorphism ranking the best across these three factors. This provides further detail to prior work by Yang and Lee, which highlights the effect of visual attractiveness on usage adoption and intention [48].

In interpreting our quantitative results, which indicate that anthropomorphism does not linearly affect users' impressions, we

draw on comments provided by our participants. The appearance of Prototype B, which received the lowest scores, was frequently critiqued; "*Fairly machine-like and inhuman, also not particularly reliable so it's not a virtual assistant that I would want to use*" (P149). In contrast, the absence of anthropomorphism in Prototype A and the high level of anthropomorphism in Prototype C positively increased participant perceptions of the prototype's abilities. For Prototype A, a participant commented on the "*very sleek and stylish*" (P67) design of the prototype. Prior work highlights how the machine-like appearance of robots provides additional confidence in their performance [49]. In contrast to these findings that highlight the benefits of machine-like appearances, we find that the version of FiPA with the highest level of anthropomorphism (Prototype C) typically outperforms the two other prototypes (see Table 2). How to interpret these at first sight contradictory results? PAs perform a distinctly different tasks from robots, and as such the effect of their appearance may differ. PAs typically perform non-critical tasks, under direct supervision of the user, with a varying degree of success in recognising user intent – as represented in our study. Given the context in which PAs are used, we hypothesise, as based on our participants' comments and prior work, that the more human-like appearance of our prototype provided a higher level of social presence to the PA [22]. This level of social presence is ultimately reflected in the higher levels of perceived intelligence, likeability, and pragmatic quality. Given the currently experienced high number of errors made by PAs when communicating with users, a 'cute' appearance may help to overcome difficulties in adoption and accurately set expectations of its performance towards future users.

5.2 Expectation Management

The increase in 'Perceived Intelligence' for the most anthropomorphic design can be considered as a positive change, yet this needs to be taken with caution when designing PAs. A study on *virtual* digital assistants, by Knijnenburg & Willemsen [16], identified that a higher degree of anthropomorphism can shape the users' mental

model of the agents capabilities, thereby leading to inflated expectations. In this study, we were able to identify similar tendencies when investigating *physically embodied* PAs. While Knijnenburg & Willemsen [16] combine multiple cue's to illustrate the system's capabilities, such as the visual appearance of the virtual assistant as well as the human likeness of verbal phrasing, we focused exclusively on embodied anthropomorphism. Knijnenburg & Willemsen highlight how this over inflation of belief in the system's capabilities was ultimately a hindrance in learning efficient interaction with the system [16]. While we did not let participants interact with the system, we do expect similar effects to occur for physical embodied PAs. We observed the same tendencies of assuming higher Perceived Intelligence, given a higher degree of anthropomorphism.

Therefore, utilising human-like features in PAs could result in increased expectations in device capabilities, leading to less optimal user experience if these expectations cannot be matched. Numerous HCI researchers have pointed out that commercially available PAs are still rudimentary in many aspects [7, 26] and that managing expectation is an important design consideration.

5.3 Physical Embodiment

This paper investigates whether it is advantageous to include anthropomorphic features in the physical design of embodied PAs. This research question corresponds to a recent trend among commercially available PAs to increasingly utilise physical anthropomorphism, for example, the Lynx [6] and – to a smaller extend – Jibo [39]. The most and least anthropomorphic prototypes (C and A respectively) scored higher in all variables than Prototype B which was moderately anthropomorphic. Our results indicate a U-shaped trend. Therefore further investigation might be relevant to identify the cause. One explanation for our results could be the widely reported uncanny valley [30] effect, which has been correlated with the U-shape in a numbers of cases including both anthropomorphism as well as zoomorphism [11, 25, 29, 38].

A study by Mathur et al. [29], investigating the effect of anthropomorphism in facial images, found a trend that is similar to our findings. They selected a comparable set (in terms of size, framing, angle etc.) of six pictures of faces (from robot - human) and asked 52 participants to rank these according to Likeability and Trust. While we did not measure trust for the three versions of FiPA, we were able to observe a similar pattern for Likeability. We found the highest values of Likeability for the prototype with the highest degree of anthropomorphism (C) followed by the least anthropomorphic (A). In a follow up replication study Mathur et al. [29] confirmed a similar trend using a different set of stimuli as well as 105 new participants. The fact that multiple studies investigating anthropomorphism and zoomorphism in a non-PA context achieve similar results to ours provides further confidence in the explanation that our observations could be related to the uncanny valley effect.

As we only have three levels of anthropomorphism in our study, we cannot report with certainty the presence of the uncanny valley in the context of PAs. This would require a follow up study involving a larger set of devices. Yet, we believe that some evidence for the presence of an uncanny valley effect are present. Adjectives used by participants in their description of the prototypes also point

towards a potential uncanny effect for Prototype B. Whereas Prototype A is described with terms of being very machine-like (*e.g.*, 'Box', 'Square', 'Machine-like') and Prototype C is associated with human-like features (*e.g.*, 'Aware', or 'Human-like'), Prototype B is the only one described with adjectives pointing towards an uncanny valley effect. Examples of terms used to describe Prototype B by our participants were: 'Scary', 'Empty look', 'Dodgy', 'Uncanny', 'Creepy', 'Kinda scary with the small 'eyes", or even 'Evil'.

5.4 Limitations and Future Work

The current study aims to collect insights into participant perceptions towards three different embodiments of a fictitious personal assistant. While our study approach, in which participants are shown videos of the prototypes in action, has proven useful in prior work, see *e.g.* [28, 42], our results can only be viewed as a study of first impressions and cannot account for long terms effect of PA design. In the case of personal assistants, this is noteworthy given the fact that a repeated misunderstandings between user and PA can quickly build up the user's frustration. We, therefore, recommend future work to explore the effect of embodied anthropomorphism over longer periods of time in the actual context of use (*i.e.*, the home) and across diverse population samples (*e.g.*, older adults [5]) and cultures.

6 CONCLUSION

In this paper we describe two studies with the goal of investigating the impact of embodied anthropomorphism on *Overall Goodness*, *Perceived Intelligence*, *Likeability*, *Hedonic*, and *Pragmatic* qualities. To this end, we developed three versions of a fictional personal assistants (FiPA). In our validation study, we asked participants to assess these prototypes – purely based on pictures – according to degree of anthropomorphism. We conclude that all three prototypes represented significantly different degrees of anthropomorphism ('low', 'moderate' and 'high'). Following this, we conducted an online between-subject study with 150 participants evaluating the three FiPA prototypes across six video showcasing typical use cases (three successful, three unsuccessful interactions). Using the widely used Godspeed and AttrakDiff questionnaires, we collected user ratings for the five dependent variables. We were able to confirm a significant increase on Perceived Intelligence, Likeability as well as the Pragmatic qualities with increase in anthropomorphism of FiPA between prototype B and C. Our findings highlight the necessity of careful consideration when designing an embodied AI (such as a personal assistant), as the physical embodiment can significantly impact people's perceptions of the device and thereby create distorted impressions of a device's capabilities.

REFERENCES

[1] Antonella Angeli, Jan Hartmann, and Alistair Sutcliffe. 2009. The Effect of Brand on the Evaluation of Websites. In *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part II* (Uppsala, Sweden) *(INTERACT '09)*. Springer-Verlag, Berlin, Heidelberg, 638–651. https://doi.org/10.1007/978-3-642-03658-3_69

[2] Jacqueline D. Bailey and Karen L. Blackmore. 2017. Gender and the Perception of Emotions in Avatars. In *Proceedings of the Australasian Computer Science Week*

Multiconference (Geelong, Australia) *(ACSW '17)*. Association for Computing Machinery, New York, NY, USA, Article 62, 8 pages. https://doi.org/10.1145/3014812.3014876

[3] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81.

[4] Jessy Ceha, Ken Jen Lee, Elizabeth Nilsen, Joslin Goh, and Edith Law. 2021. *Can a Humorous Conversational Agent Enhance Learning Experience and Outcomes?* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3411764.3445068

[5] Kyungjin Chung, Young Hoon Oh, and Da Young Ju. 2019. Elderly Users' Interaction with Conversational Agent. In *Proceedings of the 7th International Conference on Human-Agent Interaction* (Kyoto, Japan) *(HAI '19)*. Association for Computing Machinery, New York, NY, USA, 277–279. https://doi.org/10.1145/3349537.3352791

[6] Ubtech Robotics Corp. 2021. Lynx - Amazon Alexa Enabled Smart Home Robot. https://www.amazon.com/stores/page/E3365880-A1F0-402E-83E6-E83C11A0181C?ingress=0&visitId=eb6de2e6-12f7-4d5a-84c0-4a5674629794

[7] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What Can I Help You with?": Infrequent Users' Experiences of Intelligent Personal Assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 43, 12 pages. https://doi.org/10.1145/3098279.3098539

[8] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar.Help: Designing a Workflow-Based Scheduling Agent with Humans in the Loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2382–2393. https://doi.org/10.1145/3025453.3025780

[9] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & Computer 2003*. Springer, Wiesbaden, 187–196. https://doi.org/10.1007/978-3-322-80058-9_19

[10] Marc Hassenzahl and Andrew Monk. 2010. The Inference of Perceived Usability From Beauty. *Human−Computer Interaction* 25, 3 (2010), 235–260. https://doi.org/10.1080/07370024.2010.500139

[11] Chin-Chang Ho and Karl F MacDorman. 2017. Measuring the uncanny valley effect. *International Journal of Social Robotics* 9, 1 (2017), 129–139.

[12] Theodore Jensen, Yusuf Albayram, Mohammad Maifi Hasan Khan, Ross Buck, Emil Coman, and Md Abdullah Al Fahim. 2018. Initial Trustworthiness Perceptions of a Drone System Based on Performance and Process Information. In *Proceedings of the 6th International Conference on Human-Agent Interaction* (Southampton, United Kingdom) *(HAI '18)*. Association for Computing Machinery, New York, NY, USA, 229–237. https://doi.org/10.1145/3284432.3284435

[13] Kwangmin Jeong, Jihyun Sung, Hae-Sung Lee, Aram Kim, Hyemi Kim, Chanmi Park, Yuin Jeong, JeeHang Lee, and Jinwoo Kim. 2018. Fribo: A Social Networking Robot for Increasing Social Connectedness Through Sharing Daily Home Activities from Living Noise Data. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. ACM, New York, NY, USA, 114–122. https://doi.org/10.1145/3171221.3171254

[14] Michiel Joosse, Manja Lohse, Niels van Berkel, Aziez Sardar, and Vanessa Evers. 2021. Making Appearances: How Robots Should Approach People. *J. Hum.-Robot Interact.* 10, 1, Article 7 (Jan. 2021), 24 pages. https://doi.org/10.1145/3385121

[15] K. Kim, L. Boelling, S. Haesler, J. Bailenson, G. Bruder, and G. F. Welch. 2018. Does a Digital Assistant Need a Body? The Influence of Visual Embodiment and Social Behavior on the Perception of Intelligent Virtual Agents in AR. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, Manhattan, New York, U.S., 105–114. https://doi.org/10.1109/ISMAR.2018.00039

[16] Bart P. Knijnenburg and Martijn C. Willemsen. 2016. Inferring Capabilities of Intelligent Agents from Their External Traits. *ACM Trans. Interact. Intell. Syst.* 6, 4, Article 28 (Nov. 2016), 25 pages. https://doi.org/10.1145/2963106

[17] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for Workplace Reflection: A Chat and Voice-Based Conversational Agent. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 881–894. https://doi.org/10.1145/3196709.3196784

[18] Itaru Kuramoto, Jun Baba, Kohei Ogawa, Yuichiro Yoshikawa, Takayuki Kawabata, and Hiroshi Ishiguro. 2018. Conversational Agents to Suppress Customer Anger in Text-Based Customer-Support Conversations. In *Proceedings of the 6th International Conference on Human-Agent Interaction* (Southampton, United Kingdom) *(HAI '18)*. Association for Computing Machinery, New York, NY, USA, 114–121. https://doi.org/10.1145/3284432.3284457

[19] Guy Laban and Theo Araujo. 2020. Working Together with Conversational Agents: The Relationship of Perceived Cooperation with Service Performance Evaluations. In *Chatbot Research and Design*, Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie Lai-Chong Law, Ole-Christoffer Granmo, Ewa Luger, and Petter Bae Brandtzaeg (Eds.). Springer International Publishing, Cham, 215–228.

[20] Digital Dream Labs. 2021. Meet Vector 2.0. https://www.digitaldreamlabs.com/pages/meet-vector

[21] Helmut Leder, Benno Belke, Andries Oeberst, and Dorothee Augustin. 2004. A model of aesthetic appreciation and aesthetic judgments. *British journal of psychology* 95, 4 (2004), 489–508. https://doi.org/10.1348/0007126042369811

[22] Jae-Gil Lee, Ki Joon Kim, Sangwon Lee, and Dong-Hee Shin. 2015. Can Autonomous Vehicles Be Safe and Trustworthy? Effects of Appearance and Autonomy of Unmanned Driving Systems. *International Journal of Human−Computer Interaction* 31, 10 (2015), 682–691. https://doi.org/10.1080/10447318.2015.1070547

[23] L. Li, K. Ota, and M. Dong. 2018. Humanlike Driving: Empirical Decision-Making System for Autonomous Vehicles. *IEEE Transactions on Vehicular Technology* 67, 8 (2018), 6814–6823. https://doi.org/10.1109/TVT.2018.2822762

[24] Lanna Lima, Vasco Furtado, Elizabeth Furtado, and Virgilio Almeida. 2019. Empirical Analysis of Bias in Voice-Based Personal Assistants. In *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 533–538. https://doi.org/10.1145/3308560.3317597

[25] Diana Löffler, Judith Dörrenbächer, and Marc Hassenzahl. 2020. The Uncanny Valley Effect in Zoomorphic Robots: The U-Shaped Relation Between Animal Likeness and Likeability. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 261–270. https://doi.org/10.1145/3319502.3374788

[26] 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* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[27] Y. Ma, Z. Wang, H. Yang, and L. Yang. 2020. Artificial intelligence applications in the development of autonomous vehicles: a survey. *IEEE/CAA Journal of Automatica Sinica* 7, 2 (2020), 315–329. https://doi.org/10.1109/JAS.2020.1003021

[28] Wendy E. Mackay, Anne V. Ratzer, and Paul Janecek. 2000. Video Artifacts for Design: Bridging the Gap between Abstraction and Detail. In *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (New York City, New York, USA) *(DIS '00)*. Association for Computing Machinery, New York, NY, USA, 72–82. https://doi.org/10.1145/347642.347666

[29] Maya B. Mathur and David B. Reichling. 2016. Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition* 146 (2016), 22–32. https://doi.org/10.1016/j.cognition.2015.09.008

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

[31] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3173574.3173580

[32] Jeni Paay, Jesper Kjeldskov, Kathrine Maja Hansen, Tobias Jørgensen, and Katrine Leth Overgaard. 2020. Digital ethnography of home use of digital personal assistants. *Behaviour & Information Technology* 0, 0 (2020), 1–19. https://doi.org/10.1080/0144929X.2020.1834620

[33] Eleftherios Papachristos. 2019. Assessing the performance of short multi-item questionnaires in aesthetic evaluation of websites. *Behaviour & Information Technology* 38, 5 (2019), 469–485.

[34] Eleftherios Papachristos and Nikolaos Avouris. 2013. The Influence of Website Category on Aesthetic Preferences. In *Human-Computer Interaction − INTERACT 2013*, Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 445–452.

[35] Felix Richter. 2020. Smart Speaker Adoption Continues to Rise. https://www.statista.com/chart/16597/smart-speaker-ownership-in-the-united-states/

[36] Mutsuo Sano Sano, Yuka Kanemoto Kanemoto, Syogo Noda Noda, Kenzaburo Miyawaki Miyawaki, and Nami Fukutome Fukutome. 2014. A Cooking Assistant Robot Using Intuitive Onomatopoetic Expressions and Joint Attention. In *Proceedings of the Second International Conference on Human-Agent Interaction* (Tsukuba, Japan) *(HAI '14)*. Association for Computing Machinery, New York, NY, USA, 117–120. https://doi.org/10.1145/2658861.2658901

[37] Eike Schneiders, Anne Marie Kanstrup, Jesper Kjeldskov, and Mikael B. Skov. 2021. Domestic Robots and the Dream of Automation: Understanding Human Interaction and Intervention.. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, 13 pages. https://doi.org/10.1145/3411764.3445629

[38] V. Schwind, K. Leicht, S. Jäger, K. Wolf, and N. Henze. 2018. Is there an uncanny valley of virtual animals? A quantitative and qualitative investigation. *International Journal of Human-Computer Studies* 111 (2018), 49–61. https://doi.org/10.1016/j.ijhcs.2017.11.003

[39] NTT Disruption Europe SLU. 2021. Jibo. https://jibo.com
[40] Hamish Tennent, Dylan Moore, Malte Jung, and Wendy Ju. 2017. Good vibrations: How consequential sounds affect perception of robotic arms. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Manhattan, New York, U.S., 928–935. https://doi.org/10.1109/ROMAN.2017.8172414
[41] Alexandre N Tuch, Eva E Presslaber, Markus Stöcklin, Klaus Opwis, and Javier A Bargas-Avila. 2012. The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. *International journal of human-computer studies* 70, 11 (2012), 794–811. https://doi.org/10.1016/j.ijhcs.2012.06.003
[42] Niels van Berkel, Omer F. Ahmad, Danail Stoyanov, Laurence Lovat, and Ann Blandford. 2021. Designing Visual Markers for Continuous Artificial Intelligence Support: A Colonoscopy Case Study. *ACM Trans. Comput. Healthcare* 2, 1, Article 7 (Dec. 2021), 24 pages. https://doi.org/10.1145/3422156
[43] Paul Van Schaik and Jonathan Ling. 2009. The role of context in perceptions of the aesthetics of web pages over time. *International journal of human-computer studies* 67, 1 (2009), 79–89. https://doi.org/10.1016/j.ijhcs.2008.09.012
[44] Voicebot. 2020. Smart Speaker Consumer Adoption Report.

[45] Katja Wagner, Frederic Nimmermann, and Hanna Schramm-Klein. 2019. Is it human? The role of anthropomorphism as a driver for the successful acceptance of digital voice assistants. In *proceedings of the 52nd Hawaii international conference on system sciences*. SemanticScholar, Seattle, Washington, U.S., 1386–1395.
[46] Janet H. Walker, Lee Sproull, and R. Subramani. 1994. Using a Human Face in an Interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 85–91. https://doi.org/10.1145/191666.191708
[47] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963
[48] Heetae Yang and Hwansoo Lee. 2019. Understanding user behavior of virtual personal assistant devices. *Information Systems and e-Business Management* 17, 1 (2019), 65–87.
[49] Jakub Złotowski, Hidenobu Sumioka, Shuichi Nishio, Dylan F. Glas, Christoph Bartneck, and Hiroshi Ishiguro. 2016. Appearance of a Robot Affects the Impact of its Behaviour on Perceived Trustworthiness and Empathy. *Paladyn, Journal of Behavioral Robotics* 7, 1 (2016), 55–66. https://doi.org/10.1515/pjbr-2016-0005