International Journal of Social Robotics (2023) 15:855–865
https://doi.org/10.1007/s12369-023-00990-6

Audio–Visual Predictive Processing in the Perception of Humans and Robots

Busra Sarigul[1] · Burcu A. Urgen[2,3,4]

Accepted: 3 March 2023 / Published online: 5 April 2023

Abstract
Recent work in cognitive science suggests that our expectations affect visual perception. With the rise of artificial agents in human life in the last few decades, one important question is whether our expectations about non-human agents such as humanoid robots affect how we perceive them. In the present study, we addressed this question in an audio–visual context. Participants reported whether a voice embedded in a noise belonged to a human or a robot. Prior to this judgment, they were presented with a human or a robot image that served as a cue and allowed them to form an expectation about the category of the voice that would follow. This cue was either congruent or incongruent with the category of the voice. Our results show that participants were faster and more accurate when the auditory target was preceded by a congruent cue than an incongruent cue. This was true regardless of the human-likeness of the robot. Overall, these results suggest that our expectations affect how we perceive non-human agents and shed light on future work in robot design.

Keywords Prediction · Expectation violation · Human–robot interaction · Audio–visual mismatch

1 Introduction

Advances in artificial intelligence in the last few decades have introduced us to humanoid robots that we encounter everywhere ranging from classrooms to airports to shopping malls to hospitals. While their presence in our daily lives has brought a lot of excitement, how humans perceive and interact with them has become an important research topic in cognitive science. Do we perceive them differently from the way we perceive other humans? How important is it that they look or sound human or behave like humans? What are our expectations from robots? To what extent do they fulfill our expectations? These are some of the questions cognitive scientists are interested in addressing not only to be able to better understand human nature but also to be able to guide the design of robots in the future.

In his classical work, The Design of Everyday Things, Don Norman [1] provides important insights about how cognitive sciences can help in the design of artefacts including machines such as robots. According to Norman [1], the design artifacts should be adapted to the minds of their users, and this is why one needs to understand the human mind first. This implies that a collaboration between human–robot interaction and cognitive sciences is necessary. Indeed, the use of robots in well-established cognitive psychology and neuroscience paradigms in the last decade has proven useful to understand how humans respond to non-human agents as compared to their human counterparts, and what kind of principles we should follow in humanoid robot design [2–4].

One of the cognitive psychology/neuroscience paradigms that have been successfully applied in human–robot interaction is the expectation-violation paradigm [5, 6]. These paradigms have been developed to understand the nature of information processing in a variety of perceptual and cognitive tasks [7–12] and have been instrumental to come up with recent theories of human brain and cognition such as predictive coding or computation [13–15]. According to these

✉ Busra Sarigul
b.sariguel@iwm-tuebingen.de

✉ Burcu A. Urgen
burcu.urgen@bilkent.edu.tr

[1] Leibniz-Institut für Wissensmedien, Tübingen, Germany

[2] Department of Psychology, Bilkent University, Ankara, Turkey

[3] Interdisciplinary Neuroscience Program, Bilkent University, Ankara, Turkey

[4] Aysel Sabuncu Brain Research Center, National Magnetic Resonance Imaging Research Center (UMRAM), Ankara, Turkey

theories, perception is not a purely bottom-up or stimulus-driven process, rather, expectations and prior knowledge play an important role in how we perceive our environment. A growing body of empirical work in psychology and neuroscience are in line with these theories showing that participants respond faster and more accurately when they perceive events that are expected compared to the ones that are unexpected [9–12]. These results suggest that humans constantly predict what would come next and this in turn determines what they perceive [14].

Recent work at the intersection of cognitive science and social robotics has shown that humans can extend their prediction skills to the perception of robots and form expectations about robots based on their prior experience [5, 6, 16, 17]. These studies manipulated expectations towards robots by means of using stimuli that have mismatches in a variety of visual dimensions including appearance (form), motion, and interaction. In other words, these mismatch paradigms aim to induce certain expectations based on a particular cue, and at the same time present another cue that usually does not match that cue, resulting in expectation violation. For instance, Urgen et al. [6] show that the appearance of a robot can elicit certain expectations in humans about how the robot would move, and when the robot does not move in an expected way, an N400 ERP effect is observed indicating that the expectations are violated. Using a similar paradigm, [5] showed differential activity in the parietal cortex for an agent that moved in an unexpected way compared to others that moved in an expected way, which they interpreted as a prediction error within the framework of predictive coding [13, 14]. Furthermore, in a study that investigates sensorimotor signaling in human–robot interaction, [17] shows that people show lower variability in their performance when a human-like robot commits a human-like error compared to a mechanical error and that the pattern is reversed when the agent is non-human-like morphologically.

Other HRI studies explored mismatches in multisensory contexts. While vision seems to be the dominant modality in many HRI studies that investigate how humans perceive robots, [18] highlights the critical role of voice in communication and interaction with artificial agents. Accordingly, there is a growing body of research that examines the role of voice in HRI in combination with other visual features such as the appearance or movement of artificial agents [19–30]. For instance, several studies show that the mismatch between the visual appearance and voice of an artificial agent induces the uncanny valley effect [19], impairs emotion recognition, and negatively impacts likability and believability [30]. In a similar vein, [22] shows that the inconsistency between the facial proportions and vocal realism of an artificial agent reduces its credibility and attractiveness. A study with children [20] shows that the interaction between voice and other visual features such as appearance and movement affect the perceived lifelikeness and politeness of a robot. People also find artificial agents with a human-like voice more expressive, understandable, and likable [21], or attribute more human-like attributes evidenced by drawing tasks (such as facial features) [29], than the ones with a synthetic voice.

One drawback of many studies that study HRI in an audio–visual context is that they usually use subjective measures in the form of self-reports such as fear and eeriness [19], credibility or attractiveness [22], politeness and lifelikeness [20], likability, expressiveness, and understandability [21], drawings [29], or emotion labeling [30] to evaluate artificial agents rather than more objective measures such as reaction time or accuracy. Although self-reports can be instrumental in providing an initial assessment and uncovering social behavior under a variety of tasks, they fall short for a number of reasons. First, self-reports are susceptible to the awareness and the expressiveness of the participants and may provide an incomplete or biased picture of human behavior if participants lack these skills [31]. Second, self-reports usually do not provide a mechanistic understanding which would help with both explaining and predicting human behavior [6, 32]. Indeed, Greenwald and Banaji [33] recommend the use of implicit measures to better understand human social cognition. To support this effort, many tasks have been developed such as priming and implicit association tasks that usually rely on reaction times [34], as well as eye-tracking [32] and neurophysiological measures [6] recorded within strong cognitive psychology paradigms in human–robot interaction. Some studies even directly compared the results of explicit and implicit measurements. A common finding of these studies is that explicit and implicit measures are modulated differently by the experimental conditions that are under investigation [32, 35, 36]. Therefore, given the limitations of explicit measures, it is important to benefit from implicit measures recorded under well-established paradigms to gain a better understanding of human perception and cognition in human–robot interaction, especially in multisensory contexts.

The aim of the present study is to investigate the perception of human and synthetic voices in the presence of congruent or incongruent visual cues about the agents that produce those voices using a prediction paradigm. More specifically, we aim to address whether we make predictions about how robots *sound* based on how they *look* and whether those predictions are similar to the ones we make for humans. To this end, we used an expectation-violation paradigm in which human participants judged whether a greeting word sounded human-like or synthetic (‘robotic’). This sound was preceded by a picture of a human or a robot and informed the participants with a certain probability about the sound that would follow (thus form expectations). The hypothesis is that people would discriminate the robot sounds faster when they are preceded by a robot picture in contrast to a human picture just as they would do so with human sounds that are preceded by human pictures.

2 Method

2.1 Participants

30 healthy adults from the university community (16 females, Mean age = 25.2, SD = 0.65) participated in the experiment. All participants had normal or corrected-to-normal vision and hearing. The sample size of the study was determined by a power analysis prior to data collection. The minimum required sample size was determined to be 30 by using G*Power (with alpha = 0.05, beta = 0.90, $\eta^2 = 0.25$). The study was approved by the Human Research Ethics Committee of the university and all subjects signed a consent form before the study.

2.2 Stimuli

2.2.1 Visual Stimuli

The visual stimuli consisted of static images of three agents. We call them Human, Android, and Robot (see Fig. 1). Android and Robot are the same machine in different appearances. Android has a more human-like appearance, and was modeled from the Human agent, whereas Robot has a more mechanical appearance as the clothing is removed. Android is the robot Repliee Q2 which was developed at Osaka University. The images in Fig. 1 were captured from the videos of Saygin-Ishiguro database [5, 6], the agents were doing hand waving gesture. The images were 240 × 240 pixels in size, and all three were matched in terms of their low-level properties (luminance and spatial frequency) with SHINE Toolbox [37].

2.2.2 Auditory Stimuli

The auditory stimuli consisted of two sound files which lasted 2 s: the voice of a human saying ‘Good morning’ (Human Voice), and a modified version of it in which the voice sounds synthetic (we call it ‘Robotic Voice’ within the context of this study). We explored several sound programs which create synthetic voices usually associated with robots considering our experience with science-fiction movies, smart devices, voice assistants, and video games, and discovered that the main manipulation on these sounds is to play with its echo and frequency. To create a synthetic voice that would be associated with a robot in a controlled manner, we modified the human voice by means of manipulating only these two features and keeping everything else constant. We used the audio library AudioLib in Python [38] for this modification. The library has 5 different sound types: Ghost, Radio, Robotic, Echo, and Darth Vader. We conducted a pilot study in the lab with a small group of people to check whether applying any of these filters actually worked, and the ghost was found to be the most synthetic sound that was associated with a robot. To compensate for the echo factor in this synthetic voice, echo (0.05) was added to the human voice. Human and synthetic audio files were otherwise matched in terms of their amplitude (i.e., loudness) using Adobe Audition CC (13.0.6). We also added white noise to both sound files to make the task harder, as previous research on prediction shows that the effect of prediction is strongest when the stimulus is ambiguous [11]. In order to decide on the task difficulty, we added different levels of white noise (soft: 1/20, medium: 1/8, severe: 1/2) and tested them in a pilot study. It suggested that the visual cue (i.e. prior) was used when the fraction of the white noise level was 1/2.

2.3 Procedure

Subjects participated in two experiments (Experiment 1 and Experiment 2). The order of the experiments was counterbalanced across subjects. In both experiments, the subjects were seated 57 cm away from a computer screen. Their heads were stabilized with a chinrest. Before each experiment, the subjects were introduced to visual and auditory stimuli and were given verbal instructions. When introducing the visual stimuli, it was stated to the participants that the Android is a type of robot. In addition, the human voice was told to belong to the human in the Human image, and the synthetic voice was told to belong to the agent in the Robot or Android image depending on the experiment. They also did a practice session to make sure that they understood the task. The experiment was programmed in Psychtoolbox-3 [39, 40].

2.3.1 Experiment 1

Experiment 1 consisted of 5 blocks, each containing 80 trials. Each trial started with a fixation cross on a gray background (1 s), which was followed by a visual cue (1 s), an image of a human or a mechanical robot (Human and Robot agents in Fig. 1). Following the visual cue, a 2 s auditory stimulus was presented, either a human or a synthetic (robotic) voice (See Fig. 2). The task of the subjects was to indicate whether the sound was human-like or robot-like by pressing a key.

The visual cue informed the subjects about the upcoming auditory stimulus category. Following the previous work that used prediction paradigms [10, 11, 41, 42], in 80% of the trials, the visual cue was congruent with the auditory stimulus (e.g., human image and human voice or robot image and robotic voice), whereas, in 20% of the trials, the visual cue was incongruent with the auditory stimulus (e.g. human

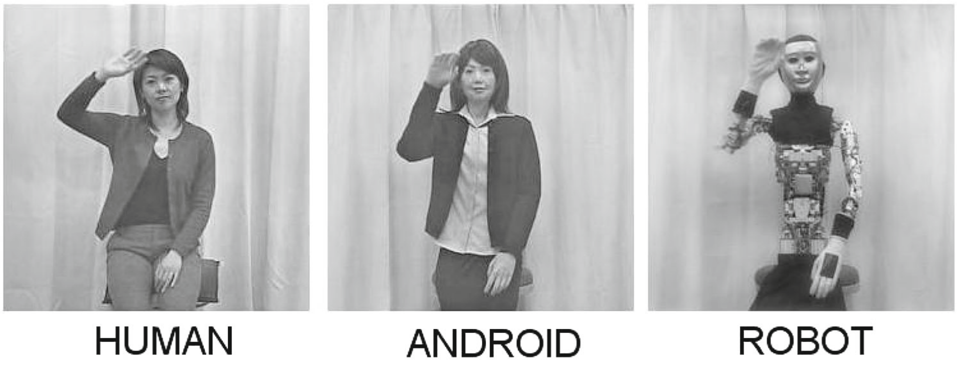

Fig. 1 Visual stimuli in the experiments consist of images of three agents with different degrees of human-likeness: a human (Human), and two robots, one having more human-like appearance (Android) and one having less human-like appearance (Robot)

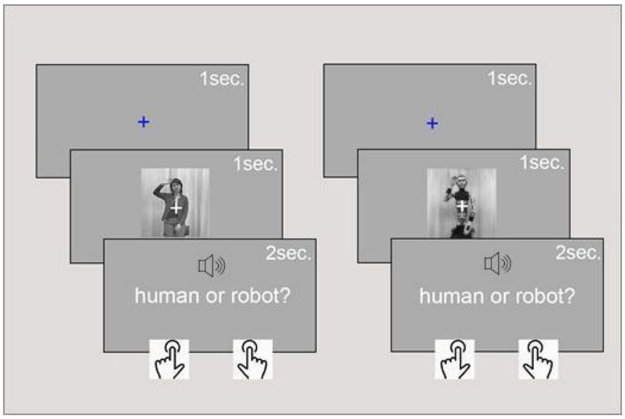

Fig. 2 Each trial in Experiment 1 consists of a fixation screen, a visual cue (Human or Robot) and an auditory target (human or robotic voice) after which subjects need to respond with a key press

image and robot-like voice or robot image and human voice, see Fig. 3).

2.3.2 Experiment 2

Experiment 2 is identical to Experiment 1 except the visual cue screen. As a visual cue, subjects were shown the image of either a human or a human-like robot (Human and Android agents in Fig. 4).

Similar to Experiment 1, there were two types of trials: congruent trials and incongruent trials. In congruent trials (80% of total trials), the category of visual cue matched the category of the auditory stimulus (e.g., human image and human voice, or a robot image and robotic voice). In incongruent trials, the category of the visual cue did not match the category of the auditory target (e.g., human appearance and robotic voice, or robotic appearance and human voice, see Fig. 5). The total number of trials was the same with Experiment 1.

Note that we did not include the human, android, and robot conditions in a single experiment as it would require the generation of three levels of voice stimuli, which may not necessarily match perfectly with the human-likeness level of the images. So, for the sake of simplicity and interpretability, and following the two-category structure of previous prediction paradigms [12, 42], we conducted two experiments in which we compared a human and a robot, and across the two experiments, we compared the effect of human-likeness of the robot.

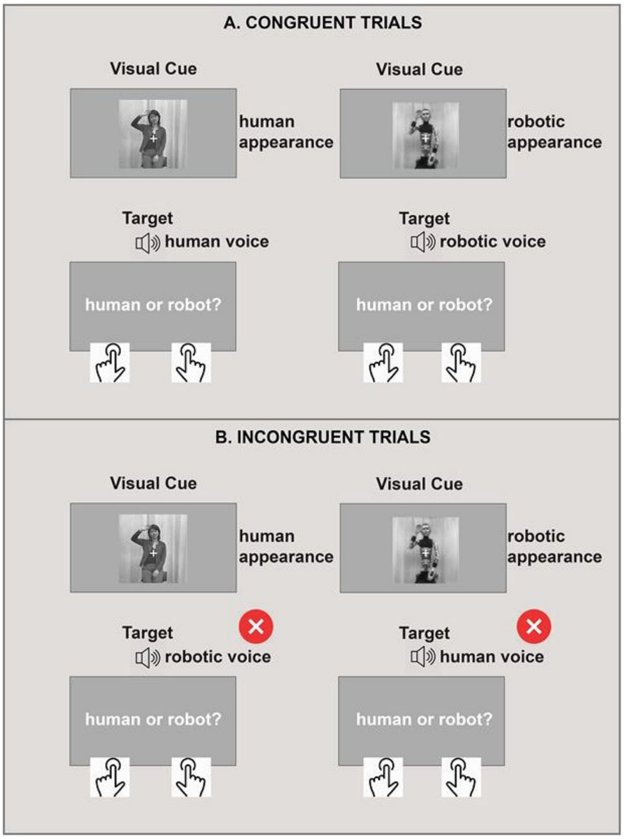

Fig. 3 There are two types of trials in Experiment 1. **A** Congruent trials in which the category of the visual cue and the auditory target match (e.g., human appearance (Human) and human voice, or robotic appearance (Robot) and synthetic (robotic) voice), **B** Incongruent trials in which the category of the visual cue and the auditory target do not match (e.g. human appearance (Human) and robotic voice, or robotic appearance (Robot) and human voice)

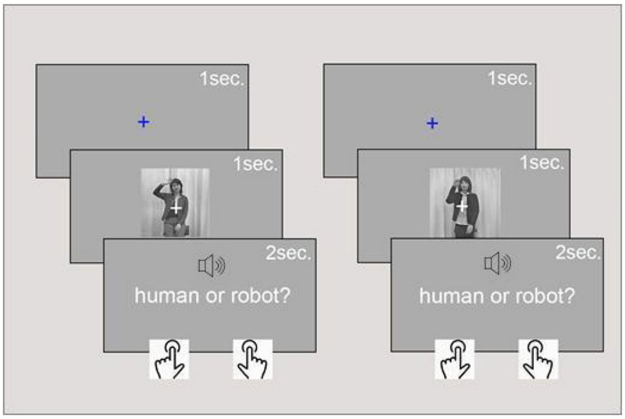

Fig. 4 Each trial in Experiment 2 consists of a fixation screen, a visual cue (Human or Android), and an auditory target (human or robotic voice) after which subjects need to respond with a keypress

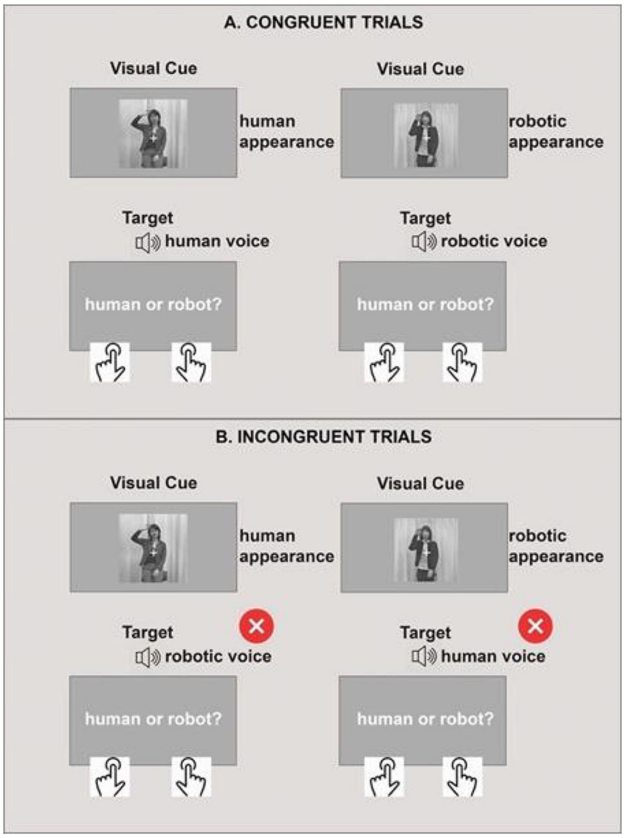

Fig. 5 There are two types of trials in Experiment 2. **A** Congruent trials in which the category of the visual cue and the auditory target match (e.g., human appearance (Human) and human voice, or robotic appearance (Android) and robotic voice), **B** Incongruent trials in which the category of the visual cue and the auditory target do not match (e.g. human appearance (Human) and robotic voice, or robotic appearance (Android) and human voice)

2.3.3 Statistical Analysis

We conducted separate ANOVAs for Experiment 1 and Experiment 2, and an additional ANOVA to compare the results of Experiment 1 and Experiment 2.

Experiment 1 We conducted 2 (Congruency: Congruent, Incongruent) × 2 (Visual Cue: Human, Robot) mixed ANOVA to investigate the effects of congruency and visual cue (agent) on reaction times and accuracy. The congruency was taken as a between-subjects factor due to the unbalanced number of trials between its levels, and the visual cue was taken as a within-subject variable.

Experiment 2 We conducted 2 (Congruency: Congruent, Incongruent) × 2 (Visual Cue: Human, Android) repeated measures ANOVA to investigate the effects of congruency and visual cue (agent) on reaction times and accuracy. The congruency was taken as a between-subjects factor due to the unbalanced number of trials between its levels, and the visual cue was taken as a within-subject variable.

Comparison of Experiment 1 and Experiment 2 We conducted a 4 (Visual Cue: Human 1 (Experiment 1), Robot, Human 2 (Experiment 2), Android) × 2 (Congruency: Congruent, Incongruent) × 2 (Experiment Order: 1, 2) mixed ANOVA to investigate whether the congruency, the human-likeness of the agent, the order of Experiment 1 and 2 (Robot or Android first) and their interaction affect reaction times or accuracy.

3 Results

3.1 Experiment 1 (Human, Robot)

3.1.1 Accuracy

The data met the assumptions of ANOVA, so we ran 2 × 2 mixed-design ANOVA with a within-subjects factor of visual cue (human, robot) and a between-subject factor of congruency (congruent, incongruent) on the accuracy scores. There was a main effect of congruency on the accuracy scores ($F(1,58) = 22.66, p < 0.05, \eta^2 = 0.28$). Congruent trials were overall more accurate than incongruent trials (Fig. 6). There was no significant effect of visual cue ($F(1,58) = 0.13, p = 0.72$), nor the interaction between congruency and visual cue ($F(1,58) = 0.09, p = 0.76$) on the accuracy scores.

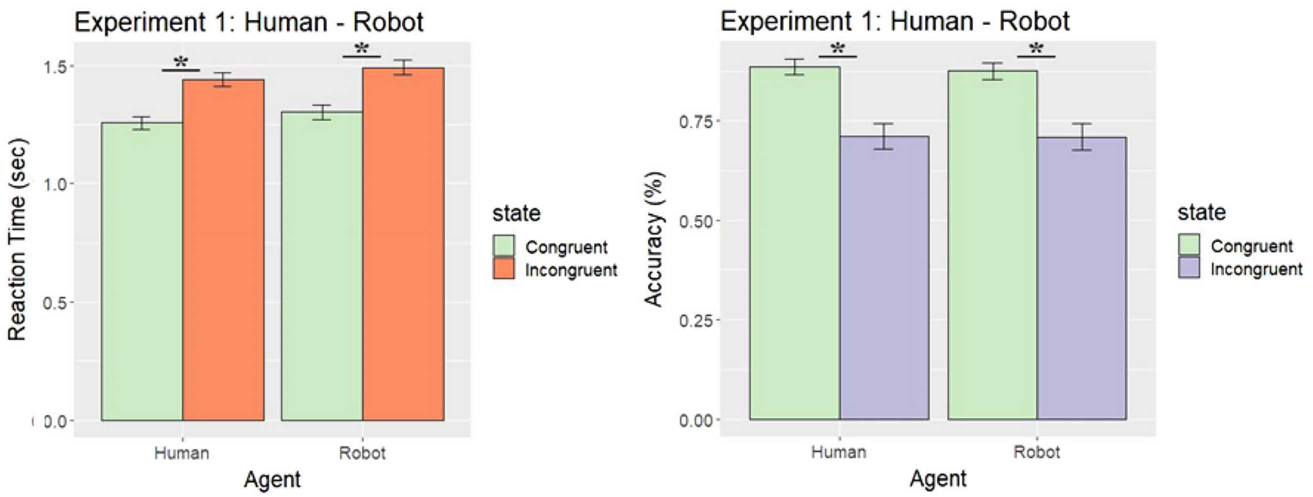

Fig. 6 Reaction Times (RT) on correct trials (left) and accuracy (%) results (right) of Experiment 1. Error bars show the standard error of the mean (SEM)

3.1.2 Reaction Times (Correct Trials)

The data met the assumptions of ANOVA, so we ran 2 × 2 mixed-design ANOVA with a within-subjects factor of visual cue (human, robot) and a between-subject factor of congruency (congruent, incongruent) on the reaction times of correct trials. There was a main effect of congruency on the reaction time of correct trials (F(1,58) = 21.27, $p < 0.05$, $\eta^2 = 0.27$). Subjects were significantly faster in congruent trials than they were in incongruent trials (Fig. 6). There was also a main effect of the visual cue on reaction times (F(1,58) = 16.20, $p < 0.05$, $\eta^2 = 0.22$). Subjects were significantly faster when the visual cue was Human than it was Robot (Fig. 6). There was no interaction between the congruency and the visual cue on the reaction times (F(1,58) = 0.04, $p = 0.83$).

3.2 Experiment 2 (Human, Android)

3.2.1 Accuracy

The data met the assumptions of ANOVA, so we ran 2 × 2 mixed-design ANOVA with a within-subjects factor of visual cue (human, android) and a between-subject factor of congruency (congruent, incongruent) on the accuracy scores. There was a main effect of congruency on accuracy scores (F(1,58) = 21.61, $p < 0.05$, $\eta^2 = 0.27$). Subjects were significantly more accurate in congruent trials than incongruent trials (Fig. 7). There was no significant effect of the visual cue on accuracy scores (F(1,58) = 0.41, $p = 0.53$). There was no interaction between congruency and accuracy either (F(1,58) = 0.04, $p = 0.85$).

3.2.2 Reaction Times (Correct Trials)

The data met the assumptions of ANOVA, so we ran 2 × 2 mixed-design ANOVA with a within-subjects factor of visual cue (human, android) and a between-subject factor of congruency (congruent, incongruent) on the reaction times of correct trials. There was a main effect of congruency on the reaction time of correct trials (F(1,58) = 20.48, $p < 0.05$, $\eta^2 = 0.26$). Subjects were significantly faster in congruent trials than they were in incongruent trials (Fig. 7). There was also a main effect of the visual cue on reaction times (F(1,58) = 25.57, $p < 0.05$, $\eta^2 = 0.31$). Subjects were significantly faster when the visual cue was Human than it was Android (Fig. 7). There was no significant interaction between congruency and visual cue (F(1,58) = 1.16, $p = 0.29$).

3.3 The Human-Likeness Dimension: The Comparison of Experiment 1 and Experiment 2

In addition to the main analyses reported above, we explored whether the human-likeness of the agent in the spectrum of Human-Android-Robot affected the reaction times or accuracy. To this end, we compared the reaction times of Experiment 1 and Experiment 2. Since we have four agents (visual cues) in the two experiments (Experiment 1: Human–Robot and Experiment 2: Human–Android), we included all of them as Human1, Robot, Human2, and Android. In addition to the visual cue, we also included the congruency and order of the experiments in a 4 × 2 × 2 mixed ANOVA.

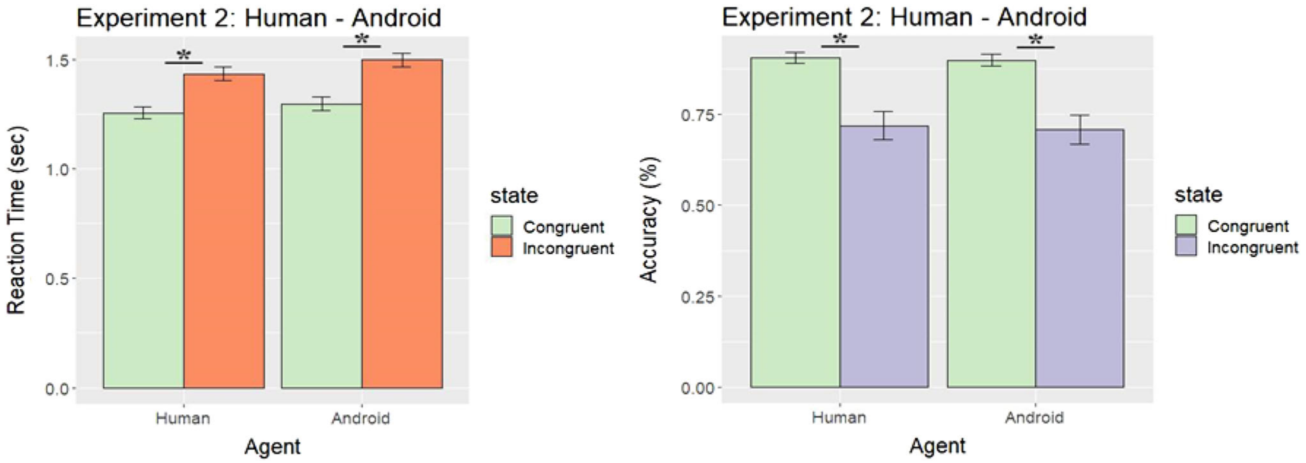

Fig. 7 Reaction Times (RT) on correct trials (left) and Accuracy (%) results (right) of Experiment 2. Error bars show the standard error of the mean (SEM)

3.3.1 Accuracy

Data met the assumptions of running an ANOVA. There was a main effect of congruency on the accuracy scores (F(1,56) = 23.93, $p < 0.05$, $\eta^2 = 0.30$). Subjects were significantly more accurate in congruent trials than incongruent trials. There was no significant effect of the visual cue on the accuracy scores (F(3,168) = 0.57, $p = 0.64$). There was no significant effect of experiment order either (F(1,56) = 0.01,$p = 0.93$). None of the interactions were significant (Cue × Congruency: F(3,168) = 0.27, $p = 0.85$; Cue × Order: F(1,168) = 0.80, $p = 0.49$; Congruency × Order: F(1,56) = 0.01,$p = 0.93$; Cue × Congruency × Order: F(3,168) = 0.07, $p = 0.98$).

3.3.2 Reaction Times (Correct Trials)

Data met the assumption of homogeneity (Levene's test $p > 0.05$) but violated the assumption of sphericity (Mauchly's test, (5) = 65.88, $p < 0.05$). Therefore, we used Greenhouse-–Geisser correction wherever needed. There was a main effect of the visual cue on the reaction times (F(1.74, 97.24) = 6.87, $p < 0.05$, $\eta^2 = 0.11$). Planned contrasts showed that the reaction times for Human 1 are significantly faster than Robot ($p = 0.01$, $\eta^2 = 0.11$) and Android ($p < 0.05$, $\eta^2 = 0.31$) but did not differ from Human 2 ($p = 0.72$); the reaction times for Robot are significantly slower than Human 2 ($p < 0.05$, $\eta^2 = 0.22$) but did not differ from Android ($p = 0.97$); and the reaction times for Human 2 are significantly faster than Android ($p < 0.05$, $\eta^2 = 0.15$).

There was a main effect of congruency on reaction times (F(1,56) = 23.36, $p < 0.05$, $\eta^2 = 0.29$). Subjects were significantly more accurate in congruent trials than incongruent trials. There was no significant effect of the experiment order on reaction times (F(1,56) = 0.01, $p = 0.96$). None of the interactions were significant (Cue × Congruency: F(1.74, 97.24) = 0.27, $p = 0.85$; Cue × Order: F(1.74, 97.24) = 0.10, $p = 0.96$; Congruency × Order: F(1,56) = 0.18, $p = 0.68$; Cue × Congruency x Order: F(1.74, 97.24) = 0.20, $p = 0.79$).

4 Discussion

We investigated whether expectations about artificial agents affect our perception. To this end, we used a well-known prediction paradigm from cognitive psychology in a human–robot interaction context. We hypothesized that people would get faster in judging how an agent sounds (human-like or synthetic) if it was preceded by a congruent visual cue (e.g. a robot picture for a robot-like voice) than an incongruent visual cue (e.g. a human picture for a synthetic, robot-like voice).

Our results suggest that people form expectations about how an agent sounds based on the visual appearance of the agent. If the visual cue is a robot, people expect that it would sound synthetic, as demonstrated by shorter reaction times and more accurate responses when the appearance and voice were congruent than when they were incongruent. This was true whether the robot has a more human-like appearance or a less human-like appearance. These results are consistent with previous work that suggests that predictive processes underlie our perception [9–11, 13, 14] including humans and robots. In other words, it seems that we can extend our predictive capabilities to perceive artificial agents, and just like our interaction with other humans, our expectations can affect how we perceive non-human agents.

An important contribution of our study to the previous work on prediction in HRI is its multimodal nature. Although there are studies that examined the effect of expectations on the perception of robots, most of these studies were done in the visual modality [5, 6, 17]. Given the recent work that highlights the importance of voice in HRI [18] and the developments in text-to-speech technology, it has become essential to go beyond the visual modality and incorporate the effects of voice on communication and interaction with artificial agents. Studies that support this effort usually manipulate the congruity of voice and appearance cues and measure a variety of things regarding the artificial agents such as their attractiveness and credibility [22], likeability and believability [21], perceived lifelikeness and politeness [20], as well as emotion recognition [30], embodiment [29], and the uncanny valley [19]. Our study extends this body of work in two ways. First, rather than presenting the visual and auditory aspects of the stimuli simultaneously, it presents them consecutively in a prediction paradigm where the visual stimulus serves as a prior (cue) for the upcoming auditory stimulus. The advantage of this method is that it allows us to study the effect of expectations on perception more directly, by involving explicit priors, rather than making assumptions or post hoc conclusions about predictive mechanisms. Second, unlike previous work that used explicit measures in the form of self-reports in the multimodal perception of robots, we used implicit measures such as reaction times and accuracy. One advantage of implicit measures is that they are more objective and less susceptible to the participants' awareness and the ability to express their introspective states [31]. More importantly, they are much better at providing a mechanistic understanding of human behavior and cognition than self-reports [31, 34], thus allowing us to make more direct links with the perception literature in cognitive sciences. Consistent with previous work on predictive processing in the perception of simple or complex object stimuli [9–11], we found that reaction times get longer, and accuracy scores get lower when we encounter artificial agents that we do not expect. This in turn suggests that our expectations affect how we perceive non-human agents as they do with other natural object categories.

Our study has several implications in various fields of HRI that intersects with predictive processing. One implication concerns the design of robots and the successful interaction between humans and robots. Previous work suggests that it is better to design artificial agents that do not violate our expectations because doing otherwise may elicit undesirable responses in humans while they interact with those agents, such as the uncanny valley [5, 6, 19, 43–48], impairments in emotion recognition [30], and decreased likeability [21] and credibility [22]. A second implication concerns the specific user groups for which the robots will be developed. Predictions stem from prior knowledge, which in turn implies that any variability in prior knowledge about robots can affect to what extent predictive mechanisms are utilized. For instance, an engineer in Japan who is heavily exposed to robots may not be surprised by a metallic-looking robot speaking with a humanlike voice, unlike a person who has never interacted with a robot. The person in the former case would generate minimal prediction errors while the latter would have large prediction errors. Similar concerns may apply when we consider different generations. For instance, children who are born in the last decade in the technology era may have different expectations from robots compared to the elderly who met robots in their adult life. Future work should investigate how familiarity with robots can affect our prediction abilities and their consequences. This will enable the design of customized robots for different end users.

Our study has several limitations. The first concerns the choice of voice stimuli. To create the synthetic, what we called 'robotic', voice, we recorded and modified a natural human voice using a variety of sound parameters (frequency and echo). We acknowledge that there is not a natural 'robotic voice' category out there. So, we did our manipulation based on what we consider how a typical robot voice sounds like based on our experience with voice assistants, smart devices, science fiction movies, and video games. We also acknowledge that not all robot voices are the same but rather there may be a family of synthetic voices that are associated with robots. An inspiration for us in creating such synthetic voices was some available software libraries that modify sound stimuli to create a variety of non-human-like sounds, e.g., ghost-like, robotic, etc. While we did not run a separate study in which we examined the discriminability of the modified voices from a natural human voice, our pilot study gave us some insights into which parameter combinations elicited the most synthetic responses. Although we found this method as the most systematic way of manipulating the voice stimuli, it has some shortcomings. First, the synthetic voice transformed from a real human voice may inherently include some human cues as compared to voices that are completely synthetic, e.g., the ones generated with text-to-speech or voice synthesis methods [49]. So, it may be difficult to categorize them as non-human. Since the robots we used as stimuli in the present study were humanoid in nature, it may not be unreasonable to have some human cues in the voice. Nevertheless, text-to-speech or voice/speech synthesis methods can be considered in future studies as an alternative. Second, it may be the way the stimuli were presented to the participants before the experiment that biased their perception of the "modified human voice" as a "robotic' voice. Future work can investigate what kind of deviations from a natural human voice would lead people to categorize voices as non-human-like or synthetic in comparison to the natural human voices and completely synthetic voices generated with text-to-speech or voice/speech synthesis methods.

A related second limitation of the study is the lack of a variety of voice stimuli that come in different degrees of human likeness, unlike visual image stimuli. Since to the best of our knowledge, this is the first study that employs an explicit prediction paradigm in a multimodal context in HRI, we wanted to keep things simple and follow similar binary paradigms in cognitive sciences [12, 41, 42]. Future work can extend this work using a variety of mismatches between voice and appearance, similar to [50].

Another line of work that was not addressed in the present study but is worth pursuing is reversing the order of the modalities in the prediction paradigm. That is, one can use the auditory stimuli as the cue (prior) and the image stimuli as the target and investigate whether voices can influence how we perceive the bodies of agents. This work could show whether the predictive processing in a multimodal context is reciprocal in nature across the two modalities involved.

5 Conclusion

We investigated whether the expectations about an agent affect how we perceive that agent. More specifically, we examined how we perceive the voice of an agent if our expectations based on what we see are not met. Our results show that the present study provides insights into how we perceive and interact with robots. It seems that we can extend our predictive capabilities to the perception of robots, and just like our interaction with other humans, our expectations can affect how we perceive robots. In sum, we would interact with artificial agents much more efficiently if they are designed in such a way that they do not violate our expectations. The use of a well-established prediction paradigm from cognitive sciences in the present study has opened a new avenue of research in human–robot interaction. Appearance and voice are only two features among many, for which we seek a match in agent perception. Future work should investigate what features of artificial agents make us form expectations, how we do that, and under what conditions these expectations are violated.

The present study sets a good example of how the collaboration between human–robot interaction and cognitive sciences can be fruitful and useful for both sides [3, 51, 52]. Our study not only suggests possible principles for robot design but also shows how fundamental cognitive mechanisms such as prediction can generalize to agents that we have not evolved with over many generations. As such, our study shows that artificial agents such as robots can be great experimental tools for cognitive science to improve our understanding of the human mind.

Acknowledgements All listed authors offered substantial contributions to the study conception and drafting of the manuscript. Furthermore, all authors approved of the final submission and agreed to be accountable for all aspects of the work. The authors would like to thank Gaye Aşkın for help with the audio stimuli.

Funding Open Access funding enabled and organized by Projekt DEAL. There is no funding information to declare.

Data Availability We provide all data and materials of this study in an Open Science Framework repository (https://osf.io/2wsug/).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics Approval Approval of the university's ethical committee was obtained before the study started (ID: 2019_01_29_01).

Consent to Participate Participants were informed about the study's procedure in the first stage and that they should only begin the experiment if they agree to the conditions described there.

References

1. Norman D (2013) The design of everyday things: revised and expanded edition. Basic books
2. MacDorman KF, Ishiguro H (2006) The uncanny advantage of using androids in cognitive and social science research. Interact Stud 7:297–337. https://doi.org/10.1075/is.7.3.03mac
3. Cross ES, Hortensius R, Wykowska A (2019) From social brains to social robots: applying neurocognitive insights to human–robot interaction. Philos Trans R Soc B 374(1771):20180024. https://doi.org/10.1098/rstb.2018.0024
4. Cross ES, Ramsey R (2021) Mind meets machine: towards a cognitive science of human–machine interactions. Trends Cogn Sci 25(3):200–212. https://doi.org/10.1016/j.tics.2020.11.009
5. Saygin AP, Chaminade T, Ishiguro H, Driver J, Frith C (2012) The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. Soc Cong Affect Neurosci 7:413–422. https://doi.org/10.1093/scan/nsr025
6. Urgen BA, Kutas M, Saygin AP (2018) Uncanny valley as a window into predictive processing in the social brain. Neuropsychologia 114:181–185. https://doi.org/10.1016/j.neuropsychologia.2018.04.027
7. Kutas M, Hillyard SA (1980) Reading senseless sentences: brain potentials reflect semantic incongruity. Science 207:203–205. https://doi.org/10.1126/science.7350657
8. Kutas M, Federmeier KD (2011) Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu Rev Psychol 62:621–647. https://doi.org/10.1146/annurev.psych.093008.131123

9. Kok P, Brouwer GJ, van Gerven MA, de Lange FP (2013) Prior expectations bias sensory representations in visual cortex. J Neurosci Res 33(41):16275–16284. https://doi.org/10.1523/jneurosci.0742-13.2013
10. Kok P, de Lange FP (2015) Predictive coding in sensory cortex. In: An introduction to model-based cognitive neuroscience, Springer, New York, pp 221–244
11. De Lange FP, Heilbron M, Kok P (2018) How do expectations shape perception? Trends Cogn Sci 22(9):764–779. https://doi.org/10.1016/j.tics.2018.06.002
12. Urgen BM, Boyaci H (2021) Unmet expectations delay sensory processes. Vis Res 181:1–9. https://doi.org/10.1016/j.visres.2020.12.004
13. Friston K (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11(2):127–138. https://doi.org/10.1038/nrn2787
14. Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci 36(3):181–204. https://doi.org/10.1017/s0140525x12000477
15. Heeger DJ (2017) Theory of cortical function. Proc Natl Acad Sci 114(8):1773–1782. https://doi.org/10.1073/pnas.1619788114
16. Ho CC, MacDorman KF, Pramono ZD (2008) Human emotion and the uncanny valley: a GLM, MDS, and Isomap analysis of robot video ratings. In: 2008 3rd ACM/IEEE international conference on human–robot interaction (HRI), IEEE, pp. 169–176. https://doi.org/10.1145/1349822.1349845
17. Ciardo F, De Tommaso D, Wykowska A (2022) Joint action with artificial agents: human-likeness in behaviour and morphology affects sensorimotor signaling and social inclusion. Comput Hum Behav 132:107237
18. Seaborn K, Miyake NP, Pennefather P, Otake-Matsuura M (2021) Voice in human–agent interaction: a survey. ACM Comput Surv (CSUR) 54(4):1–43
19. Mitchell WJ, Szerszen SKA, Lu AS, Schermerhorn PW, Scheutz M, MacDorman KF (2011) A mismatch in the human realism of face and voice produces an uncanny valley. Percept 2(1):10–12. https://doi.org/10.1068/i0415
20. Hastie H, Lohan K, Deshmukh A, Broz F, Aylett R (2017) The interaction between voice and appearance in the embodiment of a robot tutor. In: International conference on social robotics, Springer, Cham. pp 64–74. https://doi.org/10.1007/978-3-319-70022-9_7
21. Cabral JP, Cowan BR, Zibrek K, McDonnell R (2017) The influence of synthetic voice on the evaluation of a virtual character. In: INTERSPEECH, pp 229–233. https://doi.org/10.21437/Interspeech.2017-325
22. Stein JP, Ohler P (2018) Uncanny… but convincing? Inconsistency between a virtual agent's facial proportions and vocal realism reduces its credibility and attractiveness, but not its persuasive success. Interact Comput 30(6):480–491. https://doi.org/10.1093/iwc/iwy023
23. McGinn C, Torre I (2019) Can you tell the robot by the voice? An exploratory study on the role of voice in the perception of robots. In: 2019 14th ACM/IEEE international conference on human-robot interaction (HRI), IEEE, pp 211–221. https://doi.org/10.1109/HRI.2019.8673279
24. Doehrmann O, Naumer MJ (2008) Semantics and the multisensory brain: how meaning modulates processes of audio–visual integration. Brain Res 1242:136–150. https://doi.org/10.1016/j.brainres.2008.03.071
25. Hein G, Doehrmann O, Müller NG, Kaiser J, Muckli L, Naumer MJ (2007) Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. J Neurosci Res 27(30):7881–7887. https://doi.org/10.1523/jneurosci.1740-07.2007
26. Laurienti PJ, Kraft RA, Maldjian JA, Burdette JH, Wallace MT (2004) Semantic congruence is a critical factor in multisensory behavioral performance. Exp Brain Res 158(4):405–414. https://doi.org/10.1007/s00221-004-1913-2
27. Talsma D (2015) Predictive coding and multisensory integration: an attentional account of the multisensory mind. Front Integr Neurosci 9(19):19. https://doi.org/10.3389/fnint.2015.00019
28. Nie J, Park M, Marin, AL, Sundar SS (2012) Can you hold my hand? Physical warmth in human-robot interaction. In: 2012 7th ACM/IEEE international conference on human–robot interaction (HRI), IEEE, pp 201–202. https://doi.org/10.1145/2157689.2157755
29. Mara M, Schreibelmayr S, Berger F (2020) Hearing a nose? User expectations of robot appearance induced by different robot voices. In: Companion of the 2020 ACM/IEEE international conference on human–robot interaction, pp 355–356, https://doi.org/10.1145/3371382.3378285
30. Tsiourti C, Weiss A, Wac K, Vincze M (2019) Multimodal integration of emotional signals from voice, body, and context: effects of (in) congruence on emotion recognition and attitudes towards robots. Int J Soc Robot 11(4):555–573. https://doi.org/10.1007/s12369-019-00524-z
31. Nosek BA, Hawkins CB, Frazier RS (2011) Implicit social cognition: from measures to mechanisms. Trends Cogn Sci 15(4):152–159
32. Kompatsiari K, Ciardo F, De Tommaso D, Wykowska A (2019) Measuring engagement elicited by eye contact in human–robot interaction. In: 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, pp 6979–6985
33. Greenwald AG, Banaji MR (1995) Implicit social cognition: attitudes, self-esteem, and stereotypes. Psychol Rev 102(1):4
34. Fazio RH, Olson MA (2003) Implicit measures in social cognition research: their meaning and use. Annu Rev Psychol 54(1):297–327
35. Li Z, Terfurth L, Woller JP, Wiese E (2022) Mind the machines: applying implicit measures of mind perception to social robotics. In: 2022 17th ACM/IEEE international conference on human–robot interaction (HRI), IEEE, pp 236–245
36. Saltık İ (2022). Explicit and implicit measurement of mind perception in social robots through individual differences modulation, MS thesis, Bilkent University
37. Willenbockel V, Sadr J, Fiset D, Horne GO, Gosselin F, Tanaka JW (2010) Controlling low-level image properties: the SHINE toolbox. Behav Res Methods 42(3):671–684. https://doi.org/10.1167/10.7.653
38. Peirce J, Gray J, Halchenko Y, Britton D, Rokem A, Strangman G (2011) PsychoPy: a psychology software in Python. https://media.readthedocs.org/pdf/psychopy-hoechenberger/latest/psychopy-hoechenberger.pdf
39. Brainard DH, Vision S (1997) The psychophysics toolbox. Spat Vis 10(4):433–436. https://doi.org/10.1163/156856897X00357
40. Pelli DG, Vision S (1997) The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat Vis 10:437–442. https://doi.org/10.1163/156856897X00366
41. Kok P, Jehee JF, De Lange FP (2012) Less is more: expectation sharpens representations in the primary visual cortex. Neuron 75(2):265–270. https://doi.org/10.1016/j.neuron.2012.04.034
42. De Loof E, Van Opstal F, Verguts T (2016) Predictive information speeds up visual awareness in an individuation task by modulating threshold setting, not processing efficiency. Vis Res 121:104–112. https://doi.org/10.1016/j.visres.2016.03.002
43. Yamamoto K, Tanaka S, Kobayashi H, Kozima H, Hashiya K (2009) A non-humanoid robot in the "uncanny valley": experimental analysis of the reaction to behavioral contingency in 2–3 year old children. PLoS ONE 4(9):e6974. https://doi.org/10.1371/journal.pone.0006974
44. Cheetham M, Pavlovic I, Jordan N, Suter P, Jancke L (2013) Category processing and the human likeness dimension of the uncanny

valley hypothesis: eye-tracking data. Front Psychol 4:108. https://doi.org/10.3389/fnhum.2011.00126

45. Tinwell A, Grimshaw M, Williams A (2010) Uncanny behaviour in survival horror games. J Gaming Virtual Worlds 2(1):3–25. https://doi.org/10.1386/jgvw.2.1.3_1
46. MacDorman KF, Chattopadhyay D (2016) Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. Cognition 146:190–205. https://doi.org/10.1016/j.cognition.2015.09.019
47. Tinwell A, Grimshaw M, Nabi DA (2015) The effect of onset asynchrony in audio–visual speech and the Uncanny Valley in virtual characters. Int J Mech Robot 2(2):97–110. https://doi.org/10.1504/IJMRS.2015.068991
48. Lee EJ (2010) The more humanlike, the better? How speech type and users' cognitive style affect social responses to computers. Comput Hum Behav 26(4):665–672. https://doi.org/10.1016/j.chb.2010.01.003
49. Li M, Guo F, Chen J, Duffy VG (2022) Evaluating users' auditory affective preference for humanoid robot voices through neural dynamics. Int J Human–Comput Interact. https://doi.org/10.1080/10447318.2022.2108586
50. Yorgancigil E, Yildirim F, Urgen BA, Erdogan SB (2022) An exploratory analysis of the neural correlates of human–robot interactions with functional near infrared spectroscopy. Front Human Neurosci. https://doi.org/10.3389/fnhum.2022.883905
51. Saygin A, Thierry C, Urgen B, Ishiguro H (2011) Cognitive neuroscience and robotics: a mutually beneficial joining of forces. In: Robotics: science and systems (RSS)
52. Wiese E, Metta G, Wykowska A (2017) Robots as intentional agents: using neuroscientific methods to make robots appear more social. Front Psychol 8:1663

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Busra Sarigul is a research associate in Everyday Media Lab, Leibniz Institut für Wissensmedien (IWM), Germany. She is currently pursuing a doctoral degree in the Department of Psychology, University of Tübingen. She received her MS in Interdisciplinary Social Psychiatry and BA in Psychology from Ankara University. Her research interests include Human-Agent Interaction, Smart Speakers, and Multisensory Integration. She is currently working on her PhD thesis investigating the communicative qualities of human-agent interaction based on the relationship between speech styles and gender.

Burcu A. Urgen is an Assistant Professor at the Department of Psychology, Bilkent University. She is also affiliated with Aysel Sabuncu Brain Research Center and National Magnetic Resonance Research Center (UMRAM). She received her PhD in Cognitive Science from University of California, San Diego (USA) in 2015. Prior to her PhD, she did her BS in Computer Engineering at Bilkent University, and MS in Cognitive Science at Middle East Technical University. Following her PhD, she worked as a postdoctoral researcher at the Department of Neuroscience, University of Parma (Italy), with Professor Guy A. Orban. Dr. Ürgen's primary research area is human visual perception with a focus on biological motion and action perception. In addition to behavioral methods, she uses a wide range of invasive and non-invasive neuroimaging techniques including fMRI, EEG, and intracranial recordings to study the neural basis of visual perception. Her research commonly utilizes state-of-the-art computational techniques including machine learning, computer vision, and effective connectivity. Besides her basic cognitive neuroscience research, Dr. Ürgen also pursues interdisciplinary research between social robotics and cognitive neuroscience to investigate the human factors that lead to successful interaction with artificial agents such as robots. Dr. Ürgen's research is supported by TÜBİTAK and TÜSEB grants.