Giving Robots a Voice: Human-in-the-Loop Voice Creation and open-ended Labeling

Pol van Rijn
pol.van-rijn@ae.mpg.de
Max Planck Institute for Empirical Aesthetics
Frankfurt, Germany

Silvan Mertes
silvan.mertes@uni-a.de
Chair for Human-Centered Artificial Intelligence, University of Augsburg
Augsburg, Germany

Kathrin Janowski
kathrin.janowski@uni-a.de
Chair for Human-Centered Artificial Intelligence, University of Augsburg
Augsburg, Germany

Katharina Weitz
katharina.weitz@uni-a.de
Chair for Human-Centered Artificial Intelligence, University of Augsburg
Augsburg, Germany

Nori Jacoby[*]
nori.jacoby@ae.mpg.de
Max Planck Institute for Empirical Aesthetics
Frankfurt, Germany

Elisabeth André[*]
elisabeth.andre@uni-a.de
Chair for Human-Centered Artificial Intelligence, University of Augsburg
Augsburg, Germany

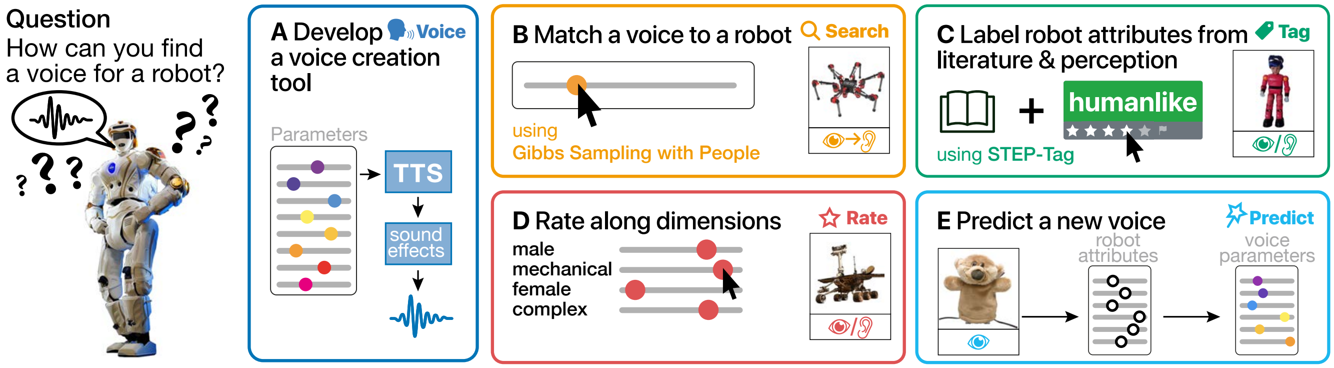


Figure 1: How can you find an appropriate voice for a robot? We propose a five-step approach: A Develop a voice creation tool for robots. B Participants iteratively change the voice of the robot using this tool to find a voice that fits well to the robot. C Identify attributes relevant to the perception of robots from previous literature and a taxonomy elicitation procedure. D A separate set of participants rates all images and their matched voices along those perceptual attributes. E Predict well-matched voices for unseen robots. For this and all future figures, the copyright holders of the robot images are defined in Tables S9–S10.

ABSTRACT

Speech is a natural interface for humans to interact with robots. Yet, aligning a robot's voice to its appearance is challenging due to the rich vocabulary of both modalities. Previous research has explored a few labels to describe robots and tested them on a limited number of robots and existing voices. Here, we develop a robot-voice creation tool followed by large-scale behavioral human experiments (N=2,505). First, participants collectively tune robotic voices to match 175 robot images using an adaptive human-in-the-loop pipeline. Then, participants describe their impression of the robot or their matched voice using another human-in-the-loop paradigm for open-ended labeling. The elicited taxonomy is then used to rate robot attributes and to predict the best voice for an unseen robot. We offer a web interface to aid engineers in customizing robot voices, demonstrating the synergy between cognitive science and machine learning for engineering tools.

CCS CONCEPTS

• **Human-centered computing → Collaborative interaction**; **Empirical studies in interaction design**; • **Information systems → Speech / audio search**; • **Computer systems organization → Robotics**.

KEYWORDS

Crowdsourcing, Personalization, Text/Speech/Language, Robot

ACM Reference Format:
Pol van Rijn, Silvan Mertes, Kathrin Janowski, Katharina Weitz, Nori Jacoby, and Elisabeth André. 2024. Giving Robots a Voice: Human-in-the-Loop Voice Creation and open-ended Labeling. In *Proceedings of the CHI*

[*] Both authors contributed equally to this research.

CHI '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642038

Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 34 pages. https://doi.org/10.1145/3613904.3642038

1 INTRODUCTION

Robots are used in a wide range of scenarios and they vary in purpose and appearance [57]. The voice is an intuitive medium for humans to interact with robots, conveying not only spoken content but also intentions [62], personality [70], conversational goals [37], and emotions [6]. However, a discrepancy between what we see (the robot's appearance) and what we hear (its voice) can strongly hinder robots' usability. Previous research has stressed the importance of users' affective responses to robots in fulfilling their functions [11, 32]. However, a mismatched voice can result in a variety of aversive reactions, such as unsettling, eerie, uncanny, and repulsive responses [51, 54, 55, 74, 78]. The intensity of this dissonance can be influenced by factors like the user's age or the robot's realism [17, 52].

Given the broad spectrum of robot designs, the range of possible voices given to robots must be similarly diverse. For example, functional robots need to be highly intelligible (e.g., for tasks like navigation), whereas popular media robots, like Pixar's *WALL-E*, are designed to sound less clear but more expressive. However, existing robot voices are limited in their diversity [12] and research is often limited in terms of the number of explored voice dimensions [31, 79, 80], the number of robots studied [2, 23, 31, 79, 80], or in terms of the robots' diversity [60]. So, how can we synthesize a robot voice, accounting for the broad spectrum of possibilities?

The advancement of Text-To-Speech (TTS) systems has made it possible to synthesize realistic human voices [77]. However, desirable robot voices may significantly differ from human voices. This work extends a state-of-the-art TTS model [35] (Figure 1A) to cover a wide range of robot voices: from highly synthetic or distorted voices to natural and individualized voices that sound similar to human speech. How can we efficiently search the expansive space of all voices to find the one that best matches a newly crafted robot? For this problem, we use an adaptive, human-in-the-loop sampling paradigm (Gibbs Sampling with People, or GSP; [24]) to iteratively find a voice that fits a robot (Figure 1B). 803 participants engaged in the task of matching a voice to 175 images of commonly used robots that span a wide variety of appearances and contexts. A separate group of human raters (N = 142) confirmed that the created voice improves over iterations and plateaus in the last iterations.

We then performed a literature review and identified attributes that characterize robots and their voices. We compared this list to labels elicited directly from participants viewing images of robots (N = 73) or listening to their matched voices (N = 59). To do so, we use a recently developed adaptive human-in-the-loop labeling paradigm [46] that does not rely on a pre-existing taxonomy (see Figure 1C). We found that terms emerging from this process mostly overlapped with those proposed in the literature. We then compiled a new list of 40 labels that frequently appear both in the literature and in our labeling pipeline and recruited new groups of participants to rate the voices (N = 245) and images (N = 298) along these dimensions (see Figure 1D). Finally, we show that the perceptual rating of the image predicts a suitable voice for a robot (see Figure 1E). A separate group of raters confirmed that the predicted voice is similarly good as the matched voice (N = 94). We conducted two separate experiments on a new set of 175 robot images from the ABOT dataset [61] (N = 249) and on randomly generated voices (N = 189) to ensure the reliability of our results. We observed that the relationship between labels in these new datasets was similar to the original one. Furthermore, using the ratings of the new robots we propose a matched voice optimized from images in the first set. We show the predicted voice is as good as the original match, confirming the robustness of our findings. We have made the developed voice creation tool publicly available as a Python package[1] to enable our validated voice configurations to be directly used in real-world applications. Finally, we provide an online robot voice prediction tool, which can be used to identify possible voices for new robots[2].

The contributions of this work can be summarized as:

- We provide a voice creation tool that covers a wide range of robotic voices using both state-of-the-art TTS and classical signal processing (Figure 1A).
- We present a human-in-the-loop approach for creating a synthetic voice for a particular robot (Figure 1B).
- We use the taxonomy elicitation process to identify labels that are relevant for the perception of robots, both in audition and vision, and compare them with attributes from the literature (Figure 1C).
- We create a densely annotated dataset of the attributes of 175 robots (Figure 1D).
- We show how our tool predicts suitable voices for new robots based on those perceptual dimensions (Figure 1E).
- In order to demonstrate that our results are robust regardless of the initial set of robots, we rerun the annotation and prediction steps with a different set of 175 robots.
- We make the resulting TTS voices publicly available as an easy-to-use software package.

2 RELATED WORK

Here we first review related research exploring the correlation between a robot's appearance and voice dimensions, underlining the significance of aligning a robot's voice with its perceived attributes. We then propose to apply two recently developed approaches to human-in-the-loop alignment to robot voice alignment: 1) human-in-the-loop sampling [24], which is the foundation of our method for collectively generating voice samples that match a robot's appearance, and 2) human-in-the-loop labeling [46], which we use to capture people's auditory or visual perceptions of robots.

2.1 Robots and Speech

Existing TTS models have frequently been used as voices for robots [68] and their quality has greatly improved in the last decade [13, 77], enabling them to produce speech that is nearly indistinguishable from human recordings [35]. humanlike voices are typically preferred over synthetic ones [40], which makes state-of-the-art TTS models an excellent voice creation tool. A recent switch from recurrent to non-autoregressive models [34, 75, 83] brought about major improvements in latency, allowing robots to produce

[1] https://robotvoice.s3.amazonaws.com/code.zip
[2] https://robotvoice.s3.amazonaws.com/predict.html

voices faster than real-time. Modern TTS models have great factorization abilities [42, 82, 89], allowing users to independently change text, prosody, and speaker identity (i.e., *what* and *how* something is being said by *whom*). Harnessing such rich latent features [72] not only facilitates the crafting of new voice personae [33, 76], but also ensures that these synthesized voices encapsulate the nuances and diversity inherent to human speech.

A substantial body of extant work emphasizes the importance of synchronizing the robot's voice with its appearance [2, 3]. Simply adopting a TTS model that delivers humanlike speech might be incongruous for a robot that has a distinctly non-human appearance. For example, imagining *R2D2* from Star Wars speaking with a plain natural voice would be odd and likely uncanny [55, 74].

So what makes a voice appropriate for a robot? Existing studies have investigated this question by looking into the correlation between appearance and voice along certain dimensions (e.g., gender or naturalness). For example, McGinn et al. [50] developed a voice association task (i.e., matching a picture of a robot to a voice) showing that gender and naturalness strongly affect the visual appearance people associate with a robot. Other studies have investigated the opposite relationship: How the voice influences the mental model that people have of a robot. For example, Powers et al. [63] showed that participants associate a male voice with a more knowledgeable person. This research also highlights the risks of reinforcing existing social biases when matching specific vocal characteristics, like a deep voice, with particular personality traits, such as being knowledgeable. In addition to aligning the voice and appearance of a robot, its behavior must also be synchronized. Torre et al. show that while trust partly depends on the voice [79], the consistency of voice and behavior is more important [80]. This is in line with previous research showing that people prefer serious-sounding robots in work-related contexts [20] and empathetic voices for healthcare robots [31].

As this literature review emphasizes, aligning robot voices with their appearances is a task of crucial importance. Here, we propose a method that can handle both highly synthetic and natural-sounding robot voices. We also develop a framework to solve the alignment problem by optimizing the voice of a specific robot based on its appearance. Finally, we enrich the aforementioned literature by providing attributes of robots that are relevant to the human perception of voices and images.

2.2 Human-in-the-Loop Sampling

Given the wide variety of possible vocal characteristics, tuning a robot's voice is a considerable challenge. While thus far this task has been left to specialists [38, 56], an alternative approach is the human-in-the-loop method. Human-in-the-loop methods efficiently integrate human decision-making with computer algorithms, so that a complex computation such as sampling or optimization can be collectively performed by humans and computers [24, 69]. Human-in-the-loop techniques have been proposed in the context of mapping internal representation in visual memory [44], 3D pose perception [43], color perception [91], musical rhythm and melody [4, 29]. More recently, various human-in-the-loop techniques have been developed in the context of speech, including human-in-the-loop evolutionary algorithms to maximize the emotional content in sound [67, 84] and a GUI-based tool allowing users to build a custom TTS voice [38]. These approaches are more efficient than elicitation methods that do not use optimization algorithms, such as reverse correlation [16, 24, 45].

A particularly efficient method for optimizing a stimulus for a desired subjective property is Gibbs Sampling with People (GSP) [24]. In this paradigm, participants are introduced to a stimulus space and use a slider interface to change one dimension of the stimulus space at a time. Importantly, the result from one iteration becomes the input for another iteration, where the same participant or a different participant now manipulates another dimension of the space (Figure 2A). For instance, when provided with a robot image, participants might adjust specific voice synthesis parameters sequentially to best align with the given image. Harrison et al. [24] demonstrated that, under experimentally verifiable conditions, this iterative method converges to samples of high subjective quality - that is, it identifies a voice that perceptually aligns with the image. GSP was previously used in the domain of emotional speech, [24, 86] or associating a particular voice to a face [85] in high-dimensional latent spaces in TTS models. To increase the speed of convergence, one can show the same slider to multiple participants and aggregate their responses (e.g., mean or median). In our experiment, we used decisions from 5 participants for every iteration.

2.3 Human-in-the-Loop Labeling

One of the best-known models for characterizing human personality is the Big Five personality taxonomy [48]. Following the Computer as Social Actors paradigm [58], researchers have applied social categories, such as gender, age, and personality, to socially interactive computer agents [32, 39]. Furthermore, specific dimensions have been developed for speech-based conversational agents [88] and for robots [8]. Despite these efforts, it remains unclear which dimensions people utilize in their perception of robots and whether these dimensions align with a range of established theories.

A recently developed adaptive tag mining pipeline called Sequential Transmission Evaluation Pipeline (STEP-Tag), where participants adaptively annotate a set of target stimuli, both by providing new descriptive tags for the stimulus and by simultaneously reviewing the tags made by previous participants [46]. When the pipeline is applied to images of robots (see Figure 2B), participants view the image of a robot, provide tags describing their impression of the robot, and rate the relevance of tags that were created by other participants (5 Likert-scale). Participants also have the possibility of flagging tags they deem inappropriate. Tags are removed if they are flagged twice (but can potentially reappear if a future participant adds them again). As the process unfolds over many iterations, meaningful tags emerge that describe the stimulus well and are validated by multiple participants, thus enabling a theory-free elicitation of tags describing the stimulus. It has previously been demonstrated that this method is effective in eliciting open-ended taxonomies without pre-specification, predicting downstream tasks (such as perceptual and semantic similarity), as well as predicting similarity in the representation of humans and deep learning models [46].

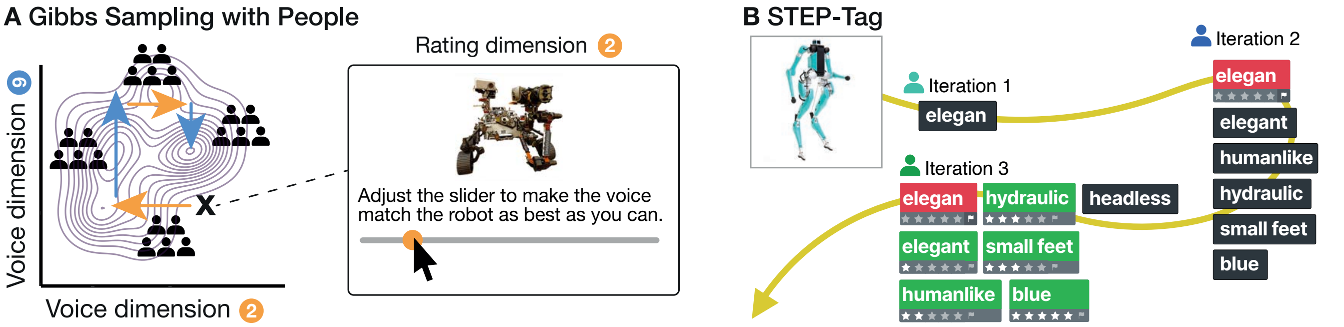


Figure 2: Human-in-the-loop paradigms. A Gibbs Sampling with People. Participants change the slider, modifying only one dimension at a time. By cycling over the dimensions, participants explore dense regions in the feature space that are associated with a given robot. B STEP-Tag. Through the labeling process, participants simultaneously create new tags and review the tags provided by others. Over many iterations, meaningful and rich semantic labels are efficiently collected for each robot image.

Unlike conventional methods in the literature, STEP-Tag eliminates the need for manual post-processing tasks like merging synonyms, thereby reducing the potential for subjectivity. However, this also means that the provided labels by the user can reflect stereotypes and reveal biases present in the data (e.g., images of cleaning robots often look feminine) and in the participants (e.g., images of masculine-looking robots are perceived as intelligent). Using STEP-Tag can help minimize prejudice in human-robot interactions by characterizing and identifying them, thus enabling engineers to create less biased systems.

3 METHODS

3.1 Images of Robots

As robots vary greatly in their appearance, our goal was to collect a variety of images that capture this variation. To simplify the complexity of possible presentation methods (such as images, videos, and 3D designs), we decided to focus on static images. We used an existing dataset (IEEE Robots) downloaded all robots from https://robots.ieee.org/robots (April 2022), removing robots without a frontal view and discarded devices such as exoskeletons or telepresence interfaces, which integrate a human user. For each robot, we selected the best image, ideally showing the entire robot in isolation. The selected images span diverse types of robots with 14 different categories, ranging from industrial to consumer robots and humanoids to drones (see Supplementary Materials C.1 for the distribution of these categories).

This list of 160 IEEE robots was extended with 15 images that were collected from other sources, such as promotional pictures from manufacturer websites or photographs taken by ourselves. To avoid contextual cues, we removed the shades and backgrounds for all robots and replaced them with a solid white background. In total, we gathered 175 images of robots across many application domains (see Table S9–S10). This selection of 175 robots is notably larger than datasets in relevant previous literature (maximally eight different robots, see Supplementary Materials C.7)

3.2 Voice Manipulation and Effects

To create a voice for a robot, we need an expressive voice creation tool that is fully parametrizable. Our solution is depicted in Figure 3. Overall, the architecture changes the voice of a Text-To-Speech (TTS) model, changes the speaking speed, and passes the resulting audio to a rack of effects. Participants use sliders to adjust the model parameters, thus changing the voice.

The first five sliders modify the voice of the speaker of a TTS model. We modified the state-of-the-art TTS model *VITS* [36] trained on the VCTK dataset [92] so that it can be used to directly modify the voice representation (speaker embedding). We performed a Principal Component Analysis (PCA) on all 110 speaker embeddings of the same dataset. We use the first five PCA components, which seem to capture sufficient variation in the human voice, as voice sliders (see Supplementary Materials C.4 for further details). These dimensions have no direct interpretation, but they correspond to intuitive vocal features such as gender, speaking speed, and voice timbre. We perform reverse PCA to obtain a speaker embedding based on the PCA dimensions. For maximum expressivity and minimum distortion, the range is constrained to approximately four standard deviations in all dimensions.

Since the variability in speaking speed in natural human speech is rather limited and the PCA dimensions by themselves did not provide enough variability in terms of duration, we added a sixth slider that can parametrically change the speaking speed ranging from 46% to 153% of the original speed using Parselmouth [30], a Python wrapper for Praat [10].

Since we could not find a suitable dataset of robot voices, we trained our TTS system on natural speech (VCTK). This means that the TTS model mainly produces naturalistic human voices and does not create robotic-like sounds. Therefore, we added sliders to apply robotic audio effects. Here we combined modern TTS with traditional signal processing techniques. We implemented eight different effects commonly used to create robotic voices: changing the pitch, decreasing synthesis quality, applying a timeshift, using a vocoder, or applying one of four different flanger configurations to the audio. We implemented an effect rack using the Librosa Python library [49], which applies the effects in a sequential order. To avoid

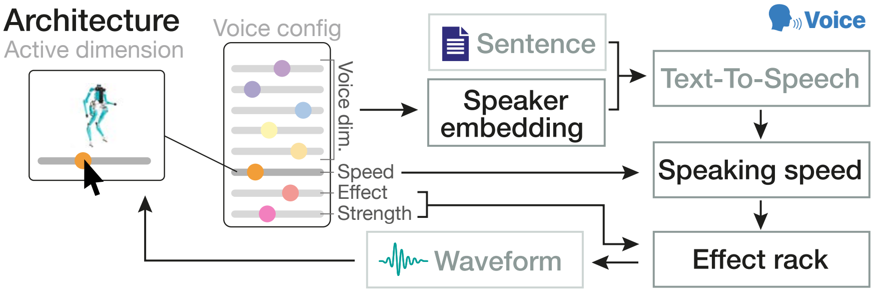


Figure 3: Architecture. The voice of the robot is controlled via eight sliders. The first five sliders control the voice of the TTS model using the first five PCA dimensions on the speaker embeddings. The sixth slider controls the speed of the speech. The seventh slider selects one of the eight effects. The last slider determines the strength of the effect. When moving the slider, the voice configuration updates one parameter in the voice configuration (here: speed). This triggers the synthesis pipeline and the resulting audio is played back to the user.

a strong mixture of voice effects, participants used a seventh slider to pick one of the eight effects and used an eighth slider to adjust the strength of the effect. The overall amplitude of the effects was manually normalized such that each effect would be approximately equally salient. The slider positions are linearly spaced (with a resolution of 16 positions) to make the synthesis computationally efficient. We used the following types of effects. Note that the exact parameters for the effects are described in the Supplementary Materials C.5 and implementation is provided in the code repository (https://robotvoice.s3.amazonaws.com/code.zip):

- **Pitch.** We enhanced the signal with two transposed audio tracks, where one was transposed five semitones up and the other transposed five semitones down. By doing so, the intonation pattern of the voice gets obscured, resulting in an unnatural voice. Further, both transposed signals are a minor seventh apart, which is generally considered a rather dissonant interval in Western music perception [14]. As such, additional tension in the voice is induced. The corresponding slider in our experiments allows us to control the ratio between the non-transposed and transposed signals.
- **Synthesis quality.** Older text-to-speech systems are poor at phase reconstruction, which results in audible artifacts that sound "robotic". To emulate this poor reconstruction, we transformed the signal to the frequency domain using a short-time Fourier Transform. We then reconstructed it using an inverse short-time Fourier Transform but with randomly initialized phase estimates. Our implementation utilized Librosa's Griffin-Lim algorithm [49] without executing phase approximation.
- **Timeshift.** To facilitate the creation of more "fuzzy" sounds, we also provided the option to blend the original voice with a slightly time-shifted version of the original signal. By doing so, the warmth and resonance of a natural voice gets veiled. To obtain this effect, the original signal was delayed for a few milliseconds, and the time-shifted signal was combined with the original signal.
- **Vocoder.** Vocoder effects are frequently used to create robotic voices [65]. We used the speech signal as a modulator for a carrier signal. By fixing that carrier signal to a certain frequency, the resulting voice sounds monotone and mechanical. Our pipeline makes use of TAL Vocoder[3], a publicly available VST implementation, which we included into our codebase using Pedalboard[4].
- **Flanger.** We incorporated a flanger, an audio effect that imparts a more synthetic quality to the sound. The flanger effect is achieved by combining a signal with a delayed version of itself where the delay time is modulated by a low-frequency oscillator. This addition offers an avenue to offset the voice's natural tone. We made four distinct flanger variants available, each producing a unique auditory experience.

For each robot we randomly selected a sentence from the 720 phonetically balanced and semantically neutral Harvard sentences [1].

3.3 Robot attributes proposed in the literature

To obtain a long list of attributes proposed for robots in literature we included labels from the "Big Five" personality model [21, 48, 66, 81], from the Godspeed questionnaires that focuses on social robots [8], from the relevant dimensions that Völkel et al. identified for voice assistants [88], and from the AttrakDiff questionnaire that focuses on user experience in general [26]. Furthermore, we added three adjectives signifying demographic features, namely "young", "male" and "female" to specify age and gender. We also added the word "animallike" because our collection of robots contains many artificial pets and animal-inspired robots. This yielded 260 unique attributes (see Supplementary Tables S4–S7 for the full list). While this list clearly does not capture all possible attributes ever mentioned for robots, it covers the most widely used attributes in the literature.

3.4 Participants and experiments

Overall, we recruited 2,505 participants. Participants were recruited from Prolific and provided informed consent under an approved

[3] https://tal-software.com/products/tal-vocoder
[4] https://github.com/spotify/pedalboard

protocol, and data was collected anonymously, with participants identified only by their prolific ID in order to enable compensation. Participants earned 9 pounds per hour, had a minimum age of 18 years, had to live and be born in the UK, had to speak English as their first language, and had to have been raised monolingually. See Supplementary Materials B.1 for additional demographic information about the participants and the number of participants in each experiment. All experiments were implemented with PsyNet [5], which is a framework for large-scale behavioral research. (see Supplementary Materials B.2) If audio was played in the experiment, we made sure participants were wearing headphones [90]. If the experiment involved a lot of text (see Supplementary Table S1), we tested English proficiency by an objective test that goes beyond their self-report (see Supplementary Materials B.3).

4 RESULTS

4.1 Human-in-the-Loop Voice Creation

803 UK participants engaged in a GSP experiment (see Supplementary Materials B.1 for demographic information). In the study, participants were tasked with tailoring voices to 175 robot images (each participant visits 20 different robots) by manipulating one slider at a time in order to tweak vocal parameters to best match the voice with the robot's appearance (as depicted in Figure 2A, see Supplementary Materials D.1 for instructions). Initially, all vocal parameters were uniformly randomized with the possible range values (see Methods, Section 3.2). We then presented the slider to five participants, and the median of their responses was carried forward to the subsequent iteration (in the case of the effects slider, we picked the majority vote, see Supplementary Materials D.2 for further justification of the choice of median). In the next iteration, the aggregated parameters from the previous generation are propagated and a new group of five participants are recruited to control a different voice dimension. The idea is that the subjective match of voices to images gradually increases over iterations [24]. The sequence in which the dimensions were altered was shuffled for each robot. The experiment concluded after 48 hours, during which time 70 images underwent 15 iterations, and 105 images experienced 16 iterations. Consequently, each of the eight dimensions was visited approximately twice.

Figure 4A shows that the standardized slider difference between consecutive iterations within a chain decreases over the course of iterations. This means that participants move the sliders to a lesser extent at later iterations, indicating convergence. In particular, there was a significant difference between the first and last iteration (Wilcoxon signed rank test: $V = 11277.0$, $n = 175$, $p < 0.001$, $r = .43$, this and all future tests are Bonferroni corrected for multiple comparisons) but we did not find a significant difference between the last iterations to the six iterations preceding it. The slider difference drops after all eight dimensions have been visited once, which is in line with previous studies [24, 85, 86]. The development over iterations can be listened to online: https://robotvoice.s3.amazonaws.com/iterations.html.

To visualize the proximity of the matched robot voices to each other, we performed a PCA on the standardized slider positions of all stimuli in the experiment. Figure 4B depicts the first two principal components and shows the distribution of all slider configurations using a kernel density estimate (gray lines). The initial robot voice configurations are uniformly sampled from the sliders but occupy distinct slider positions at the end of the experiment. For example, the spider-like robots in the upper right corner or the toy-like robots in the top left corner of the plot group together in slider space (i.e., they received similar voices in the final iteration). The final voices can be explored interactively using the online visualization: https://robotvoice.s3.amazonaws.com/explore.html.

In order to validate whether the voice and robot match improves over time, we recruited a separate group of participants ($N = 142$) that rated how well the voice matches the robot (see Supplementary Materials D.3). This experiment comprised 2,730 stimuli. All stimuli were generated in the GSP process with three additional random voices per robot. There were about 4.9 average ratings per stimulus. Overall, we had 13,597 human judgments in this experiment. As depicted in Figure 4C, the average match increases over the course of iterations. In particular, the average of the last three iterations was significantly larger than the first three iterations (Wilcoxon signed rank test: $V = 1813.5$, $n = 175$, $p < 0.001$, $r = .64$). In addition, the increase in rating over iterations reduces after each dimension is visited approximately once. For example, we did not find a significant difference between the average of iterations 8-10 and the average of iterations 13-15 (Wilcoxon signed rank test: $V = 6321.5$, $n = 175$, $p = 0.018$, $r = .10$).

4.2 Open-ended Labeling

What are the semantic labels that determine robot appearance and voice characteristics? To answer this we used STEP-Tag [46], a recently developed elicitation method to elicit labels from stimuli. We recruited two new groups of participants to annotate the obtained final robot voices and the original images ($N = 59$ and $N = 73$ respectively). Each robot is sequentially annotated by 10 participants (see Supplementary Materials E.1 for the participants' instructions). The process is adaptive: one participant provides annotation and subsequent participants rate it, flag it (in case they think it is inappropriate), or suggest their own annotation (Figure 2B). To facilitate convergence and avoid spelling variants and duplicate tags, participants can see words that start with the same letters while typing and can select them if they find them appropriate. The proposed words are either tags provided by other participants or the 260 dimensions proposed in the aforementioned literature (see Methods, Section 3.3).

As depicted in Figure 5A, the vocabulary used to describe the 175 images of robots is generally larger than those of 175 voices (765 and 217 unique tags for the image and voice modalities, respectively). Also, the same labels are used more frequently for the voice compared to the image modality (mean occurrence of 5.4 and 2.5 for the voice and image modalities, respectively).

To investigate which terms are particularly relevant, we visualize the co-occurrence network for each modality in Figure 5B using a network analysis [46]. The nodes are tags proposed in the STEP-Tag process and the edges indicate if they co-occur within the same robot with other tags. Those tags that have many connections to other tags – indicated by the larger dots – are likely to

[5] Psynet is available here: https://www.psynet.dev/. PsyNet [24] relies on the open-source platform Dallinger (https://dallinger.readthedocs.io/)

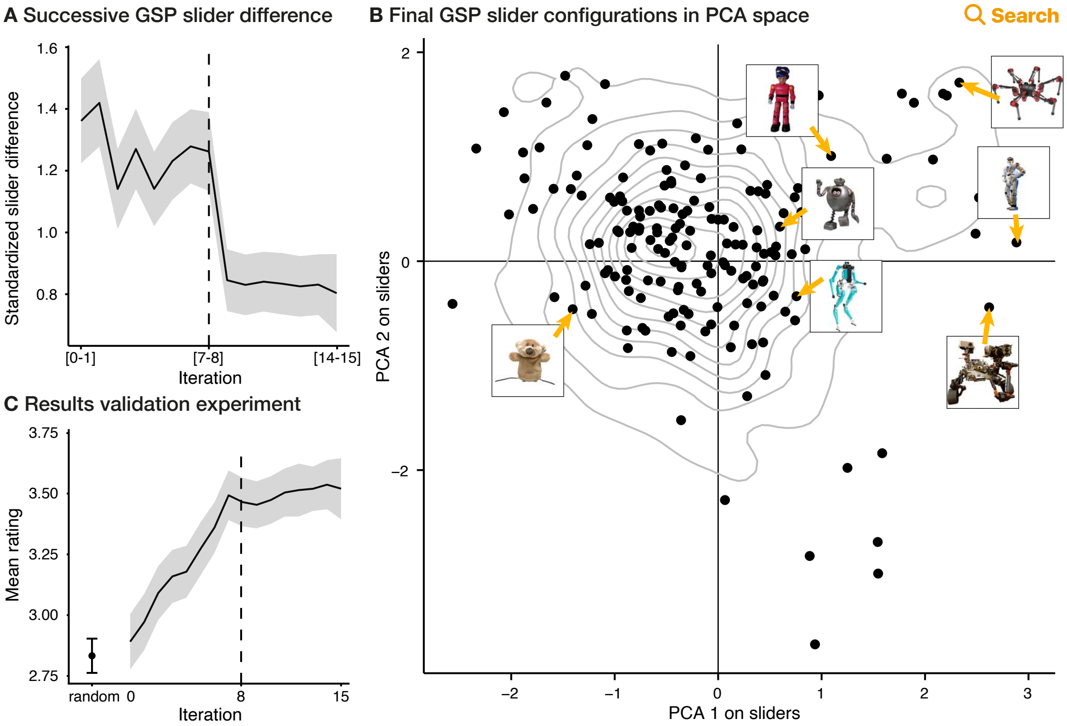


Figure 4: GSP results. A Standardized difference between successive slider configurations. B PCA on all slider configurations from all iterations. The gray kernel density estimate indicates the distribution of all slider configurations in PCA space. The black points are the final slider configurations. C Mean ratings as a function of the iterations and a random voice. Shaded areas are confidence intervals.

be relevant descriptors. In the co-occurrence network, terms that are semantically similar are often located near each other, such as 'animallike' and 'doglike' in the image modality. However, this isn't always the case, as terms that are semantically related do not necessarily appear together if they are inapplicable to a significant number of robots.

Overall, we observed that tags in the voice modality are more interconnected (average degree: image = 3.8, voice = 11.3), suggesting that a relatively small number of recurring labels frequently appear together. This observation aligns with what is shown in Figure 5A. This pattern can be partially attributed to the challenge of identifying vocal properties compared to image attributes. Voice representations might be less easily described in words, or more ambiguous overall, leading to greater overlap in semantic labels.

Interestingly, while our approach is open-ended (e.g., we don't use post-processing and involve lay participants), many central terms overlap with those commonly mentioned in literature such as "friendly", "humanlike" or "female" (see, for example, [19] or [63]). Figure 5B furthermore reveals that while some impressions are modality-specific (e.g. "high-pitched", "echo", "accent"), the majority of terms proposed by the participants reflect general impressions of the robot (e.g., "weird", "cute", "robotic" and "friendly") and are not modality-specific. However, other features differ across the two modalities. For example, the biological sex or the age of the speaker is an important category in voices, whereas for the distinction between "animallike", "humanlike", and "robotic"/"mechanical" seems more important in the case of images. Furthermore, the participants came up with terms for the voices that refer to communication qualities, such as "inaudible" or "informative", and the communication style, such as "assertive" and "unenthusiastic". Obviously, the participants were able to produce voices that complemented the visual impression of the robots by assigning additional attributes to them via the voice modality. The observation that participants used terms related to communication qualities and styles when judging voices, but not when viewing static images of robots, highlights the complementary of different sensory modalities, such as visual and auditory.

4.3 Rate Robots along Perceptual Dimensions

To understand how the perceptual dimensions in the literature and the one from the STEP-Tag procedure describe each of the robots, we performed another experiment. Here a new set of 543 participants was recruited to rate all robots across a select set of dimensions. As a result, we chose familiar dimensions in order to capture labels that existed in the literature as well as labels that were perceptually salient to participants. We selected 40 attributes in order to have enough ratings per participant. Specifically, we selected the 26 dimensions that overlap between the list of 260 labels from previous literature (see Supplementary Table S4–S7)

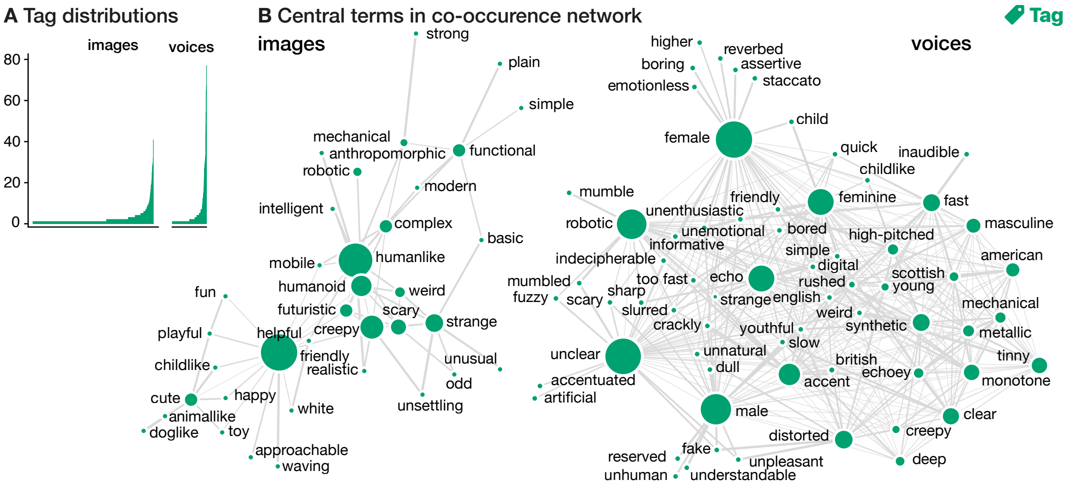


Figure 5: STEP-Tag results. A Raw occurrence of single labels for the 175 images and 175 voices. B Co-occurrence networks between provided tags per modality. Tags with a co-occurrence below 4 are pruned to remove words that are rarely used. The size of the nodes indicates the degree. Networks are created using Gephi [9].

and our STEP-Tag results. The remaining 14 dimensions are the 7 perceptually most salient features (based on STEP-Tag) in each of the modalities. Supplementary Table S8 specify the 40 dimensions and their sources.

We recruited separate groups of participants to rate these 40 perceptual labels in the image and voice modalities (N = 298 and N = 245 respectively, see Supplementary Materials F.1). Each participant rated the robot image or robot voice on 5 randomly selected dimensions using sliders that snap to 5 positions. On average, each stimulus and dimension was rated 7.5 times for the images and 6.1 times for the voices (see Supplementary Materials F.2 for the experiments' instructions). Overall, the ratings were reliable for both experiments: the split-half reliability for images was r = 0.65 and r = 0.61 for the voices. To compare the consistency of the dense rating results with the STEP-Tag results in the previous experiments, we correlated STEP-Tag ratings with the dense ratings for the labels that occur in both datasets. As shown in Supplementary Materials F.3, there is a diagonal for most terms indicating that there is a strong correlation between the number of stars a label received in STEP-Tag and the average rating it received in the dense rating experiment (mean diagonal: r = .31 and r = .24, off-diagonal: r = 0.11 and r = 0.10 for images and voices respectively).

Figure 6A shows the correlations between the dimensions for the image modality (i.e., a correlation between average rating per stimulus between all dimensions). Generally, terms with similar meanings, such as "female" and "feminine", show strong positive correlations, while antonyms like "clear" and "unclear" display strong negative correlations. The matrix reveals an additional pattern: participants tend to associate female robots with labels like "young", "playful", "cute", and "friendly", while male robots are linked with traits such as "assertive", "functional", "complex", and "intelligent". These observations align with previous literature [63], which suggests that societal stereotypes influence how robots are perceived.

For the voice, the correlation matrix shows a more consistent structure (Figure 6B): The largest cluster contains dimensions like "creepy", "unpleasant", "mechanical", and "robotic". Also "female" is associated with a "young" and "cute" voice (consistent with previous literature) [12], but not with a "friendly" voice. Instead, a new cluster emerges for "friendly", "helpful", "clear", and "intelligent" voices. This suggests that voice modality presents a much harder challenge in terms of providing labels. Specifically, voices cluster to a smaller number of interconnected terms (consistent also with the usage of smaller vocabulary in Figure 5A).

To investigate the robustness of our findings, we run the dense rating experiment on 175 new images from the ABOT dataset [61] (see Supplementary Materials C.2) and on 175 randomly created voices using our voice tool. We found strong correlations between the two image (r = .85) and two voice datasets (r = .91, see Supplementary Materials F.5). These findings indicate that the obtained correlations across the terms are robust across datasets.

We also investigated the correlations of the dimensions across the modalities. Generally, the correlations were lower, indicating that the association between the dimensions across the modalities is weaker (e.g., a masculine robot does not necessarily become a male-sounding voice, see Figure 6C for the correlation between terms across modalities). Furthermore, as depicted in Figure 6D, the diagonals between the dimensions were much weaker or entirely vanished for certain dimensions for example for terms like "humanoid" or "unpleasant"(see Supplementary Materials F.4 for a correlation matrix sorted by the strength of the diagonal). This indicates that the same labels are not consistently used across modalities, e.g. a "fast" voice does not mean that the image of the robot looks "fast" too. The dimensions that are best preserved across modalities are dimensions like "feminine", "young", and "cute" (Figure 6C).

In Figure 6D, we can see that there is a large overlap between associations from images to voices as well as from voices to images (e.g., male voices are associated with mechanical robots, and vice

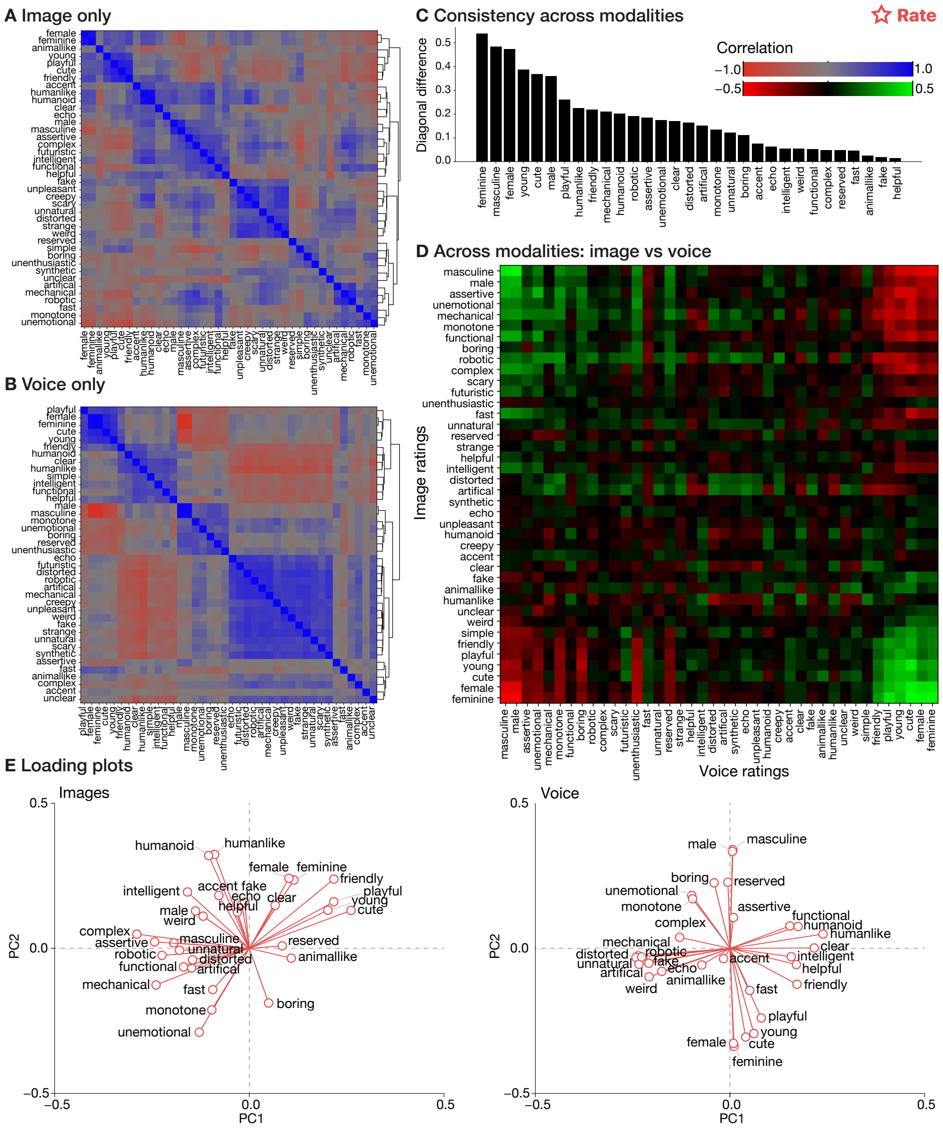


Figure 6: Correlations between ratings along dimensions. Correlation across dimensions for A images and B matched voices. Correlation matrices are sorted by the order in the dendrogram obtained via agglomerative clustering. C Most consistently rated dimensions across both modalities. The diagonal difference is the difference in correlation between the diagonal and the mean correlation of the rest of the row. D Correlation across both modalities. The correlation matrix is sorted by mean correlation for the most consistently rated dimension “feminine”. E Loading plots for both modalities. PCA components were obtained separately for the data of the correlation matrices in panels A (left) and B (right).

versa). However, this relationship is not always bidirectional; for example, assertive robots are associated with male voices (r = 0.32), but male robots are not really associated with assertive voices (r = 0.09). Further comparisons between the modalities can be made via the interactive visualization: https://robotvoice.s3.amazonaws.com/compare.html.

Figure 6E displays the factor loadings for consistent dimensions across modalities as they relate to the first two principal components in the data from the correlation matrices shown in Figures 6A and 6B. A high loading indicates a strong alignment between a specific word and the PCA dimensions. In the voice modality, the first principal component primarily captures the contrast between “humanlike” and “robotic”, while the second dimension focuses on the male-female dichotomy. In the image modality, a similar contrasting pattern is observed between “humanlike” and “robotic” features, but here the emphasis is on terms related to automaticity, such as “fast”, “monotone”, and “unemotional”, as opposed to terms like “playful”, “friendly”, and “cute”. The gender dichotomy is somewhat less pronounced here.

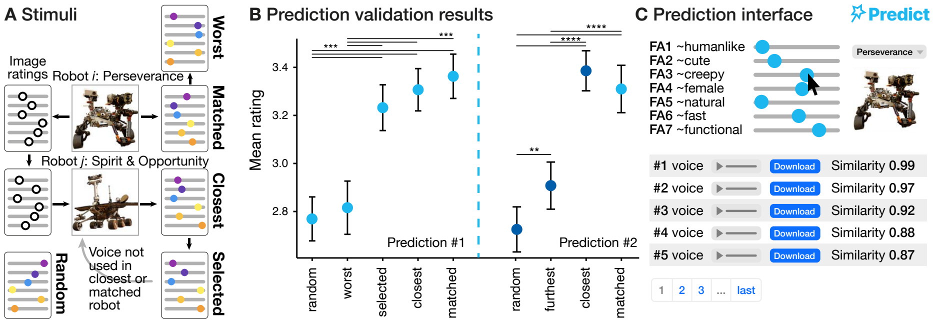


Figure 7: Prediction of a voice. A Schematics of the procedure to select stimuli for the prediction validation experiment. B Results of the prediction validation experiment. Matched receives the highest score followed by the closest and selected voice configuration. The worst and random voices receive the lowest scores and are significantly lower than the matched, closest, and selected voice configurations. * indicates that the paired Wilcoxon signed rank test was significant ($p < 0.001$). C Schematics of the prediction interface.**

4.4 Predict Voices based on Labels

Finally, we wanted to learn if our results can be used in practice for engineers who want to fit a voice to a robot. To test whether we can predict the voice of a robot based on the image ratings in the dense rating experiment we recruited a new group of (N = 94) participants. We assessed whether the obtained perceptual dimensions can be used to propose a well-matched voice to an unseen robot. For each robot image i, we provide five different combinations of an image and a voice (Figure 7A): As ground truth we included the original matched voice of robot i (*matched*). To see how well we can use verbal descriptors to perform voice prediction, we searched for the robot i with perceptual image rating across the 40 dimensions and found the closest robot j (*closest*, i.e., with the highest cosine similarity, e.g., the Perseverance robot is closest to the Spirit & Opportunity robot). We then used the voice of the j in the final iteration of the GSP experiment. To test robustness, we also searched for robot j for the GSP slider configuration and selected the closest voice in slider space, which did not occur in any iteration for robot i and j (*selected*). As a negative reference, based on the slider configuration of the matched robot i we searched for the worst slider combination (*worst*, i.e., which is maximally dissimilar in cosine similarity). Finally, we also included a random voice configuration (*random*). The interface of the prediction experiment was identical to the GSP validation experiment (see Supplementary Materials G.1). We had 875 stimuli and 7,444 human judgments overall, and each stimulus received an average of 8.5 ratings.

Consistent with the validation of the GSP voices, the random voice received the lowest voice match score and the final voice the highest match score (Figure 7B, left panel). While the closest and selected voices received a slightly lower match rating, we did not find a significant difference there (Wilcoxon signed rank test: V = 6879.0, n = 175, p = 0.47, r = .07). However, the matched, closest, and selected voices were all significantly better than the worst or random voices (p < 0.001 in all cases), which both have much lower ratings. Thus, this shows that while the predicted voices (closest and selected) were all better matches than a random voice, they were not significantly worse than the matched voice. This trend is not only visible when averaging over all participants, but also on a single-participant basis (see Supplementary Materials G.2).

To assess if our findings also extrapolate to other datasets of robots, we run another prediction experiment (N = 73). We wonder if the annotated features of the new robot can be used to match the voice based on the old data set's annotated features. In a real-world scenario where an engineer might have a new, unseen robot image and want to use our results for voice matching, this validation is crucial as it should show that even when using voices tailored to the old dataset and a matching model trained solely on the old dataset, we can still achieve accurate predictions with a new set of independently annotated images. Thus, we looked up the closest robot in terms of its annotated features for each of the new 175 robots in the old set of matched robots (*closest*). As a reference, we also included the same matched voice and paired it with the directly matched old robot image (*matched*). As a negative reference, we looked up the perceptually furthest robot in the old dataset and selected its voice (*furthest*). We also add a random voice (*random*). As shown in Figure 7B (right panel), the closest and matched voices are all significantly better than the furthest and random voice (p < 0.001 in all cases). While the closest voice received a slightly higher rating than the matched voice, this difference was not significant (Wilcoxon signed rank test: V = 8215.5, n = 175, p = 0.14, r = .11). As in the previous prediction experiment, the furthest matched voice was slightly higher than random though both of them had low ratings overall. This is probably because random voices are uniformly sampled along the dimensions, leading in some cases to sample extreme values, which is not the case for the furthest or worst voices that were matched to a robot. This additional prediction experiment shows that our prediction also works for newly annotated robots from different datasets.

To facilitate a wide adaptation of the tool, we provide an interactive voice prediction tool online: https://robotvoice.s3.amazonaws.com/predict.html (Figure 7C). The tool allows one to select a robot from the 175 robots which is most similar to the robot that requires a new voice. The user can either search for the visually closest robot from the dropdown list or modify latent dimensions representing the 40 dimensions in vision (see Supplementary Materials G.3). For example, slightly modifying the latent dimensions of the Zeno robot likely returns the voice created for the Milo robot because they look much alike. For the visual match, it will show the closest voices in the slider space. The voice configurations can be downloaded and can directly be integrated into applications.

4.5 Control analysis

In this paper, we have shown that by using human-in-the-loop approaches, participants develop voices matched to robots, identify attributes relevant to the perception of robots, and provide ratings along those dimensions that can be used to predict well-matched voices to entirely novel robots. A natural question that arises is whether neural networks can replace parts of the human pipeline. Here, we conduct two experiments on CLIP [64] (see Supplementary Materials F.8) to avoid the human dense rating experiments since they involve many participants and thus are costly. We show that there is a moderate correlation between the cosine similarity across robot images computed on the human dense rating and the image embeddings (r = .58 for the old 175 robots and r = .51 for the new robots). This indicates that CLIP embeddings provide a fair proxy for the perceived similarity of robots. In a second analysis, we use CLIP to do the dense rating experiment (so each image receives probabilities of all 40 labels). We now compute the correlation across terms for the CLIP and human data. We find a similarly strong correlation across the CLIP results across datasets (r = .87) compared to the human results (r = .85, see Supplementary Materials F.5), but the CLIP and human data are uncorrelated in both the old (r = .04) and new image dataset (r = .01). The results show that the correlational structure across the terms is consistent across the datasets, but varies greatly between CLIP and the human dense rating. While CLIP provides a proxy for the perceived similarity of robots, researchers and engineers should be cautious about blindly replacing parts of the pipeline by neural networks, as the models will introduce new biases in the annotation process.

To explore if we can actively reduce biases in our data, we rerun the STEP and dense rating experiment on the images but participants first take an implicit bias training (see Supplementary Materials C.8). To measure implicit biases before and after the training, we use the widely-used implicit association test (IAT) [22, 73]. Based on previous literature, we expected that the training would have a short-term effect on participants' responses and reduce adverse stereotypes [41]. We found that participants carefully read the implicit bias fact sheet (6/8 text-comprehension questions were answered correctly), but we did not measure a significant difference in bias before and after training.

Running the STEP experiment (N = 78), we found a large overlap in the used tags across STEP experiments with and without the training (see Supplementary Materials E.2), and the frequency of the shared tags is strongly correlated (r = .78). Moreover we found that if the data from the STEP experiment after the awareness training had been used to compile the list of 40 terms, only one term would have been replaced. These findings suggest that the STEP tag results remained largely unaffected by the training. When running the dense rating experiment (N = 202) (see Supplementary Materials F.7), we found that the correlation matrices in the dense rating experiment with and without training strongly correlate with each other (r = .91). This indicates that while participants were aware of their implicit biases (see comprehension questions), they did not substantially change their responses.

5 DISCUSSION

The present work provides a voice creation tool that can cover a wide range of robotic voices (Figure 1A). We used this tool in a human-in-the-loop approach (GSP) to create matched voices for 175 robots (Figure 1B), obtained a taxonomy using a human-in-the-loop open-ended labeling approach (STEP-Tag, Figure 1C), densely rated the attributes from the taxonomy for 175 robots (Figure 1D), and predicted suitable voices for new robots based on those perceptual dimensions (Figure 1E).

5.1 Limitation and Future work

Our paper primarily focused on conveying robot characteristics through manipulating the audio channel. To control for voice manipulation, participants were presented with static images. The way robots move can significantly affect human perception, and a wide range of literature illustrates how robots convey personality through body language, gestures, and facial expressions, as summarized in [32]. Future research can investigate how different use cases and scenarios of the same robot can affect the perceived appropriate voices. Another limitation is that the voices we used were matched with short, semantically neutral sentences, which might not generalize to longer textual content. Consequently, participants formed opinions based on limited information about the robot and its voice. An intriguing future direction for this research could include the employment of dynamic materials such as videos instead of static images, as well as the use of longer, semantically relevant spoken content. While such complexities are beyond the purview of our current study, which is focused on the vocal channel, our methodologies could serve as a foundation for more comprehensive studies into voice interactions in dynamic robot settings.

While we purposefully selected a neutral background for the robot to minimize contextual biases, it is essential to recognize that participants may have held varying perceptions of the robot's role, task, and target audience while adjusting voice dimensions. Empirical evidence indicates that factors beyond the robot's attributes, such as the task and user characteristics, significantly influence how it is perceived and how humans interact with it [39]. The transition from a toy-like robot to a robot serving as a speech assistant, as highlighted by Aylett [5] using the example of the Cosmo robot, can result in a mismatch between the robot's function and its voice. Tags used by our participants to describe the robot, such as "functional" and "helpful", highlight the importance of its intended purposes and audience in addition to its audio-visual characteristics. Both the robot's visual appearance and the mental models it triggered in participants may have influenced voice modifications. To gain further insights into these factors, conducting additional experiments with systematic changes to the robot's visual context, aligned with its intended functions, could be valuable.

At the end of the Results section, we have alluded to the possibility of replacing a part of the human pipeline with neural networks. Here, we used CLIP [64], but in future research, it would be particularly interesting to do similar experiments on Large Language

Models with vision grounding, such as Gemini, GPT-4V, or Bard, as they both have access to image and language data.

We based our voice creation tools on an English dataset, which, while diverse in including multiple dialects, did not allow us to explore the intricate relationship between culture and robot perception. This limitation applies to the user's cultural background and the culture the robot is intended to portray. Prior research [23] demonstrated that a robot's social category membership, including culture, significantly influences how people perceive and interact with it. During the annotation process, our participants included tags related to English dialects like "Scottish" and "American", highlighting the relevance of group membership as a distinguishing characteristic. McGinn and Torre [50] manipulated the accent of a robot's voice to investigate its impact on the formation of stereotypes. However, due to the heterogeneous background of the participants, their findings on the effect of accent manipulation were not consistent. Further research should, therefore, focus on the alleged cultural background of robots portrayed by their accent.

In a broader perspective, our study only involved monolingual English UK participants and future research should incorporate less "WEIRD" participants (Western, Educated, Industrial, Rich, Democratic) [7, 27] to uncover associations across perceptual dimensions in different cultures. Our approach is largely language-agnostic (it would solely require a TTS model trained on a different language) and thus can be applied to a variety of languages and cultures.

Finally, while the matched voices are significantly better than a random voice (Figure 7B), the mean ratings for the matched voices (3.4) are not quite at ceiling performance (5.0). This can have multiple causes. One explanation is that the voice model is not expressive enough yet. The voice dimensions mainly capture aspects of the voice such as gender or sex (see Supplementary Materials F.6). Future research can improve the parametrization of the latent voice dimensions to capture more expressive features of the voice. Another possible cause is that participants disagree about the voice properties associated with a robot. The split-half reliabilities in the rating experiments are high but there is still some disagreement across participants (dense: image $r = .68$ and .53, voice $r = .65$ and .48; prediction: $r = .56$ and .53). This indicates that future research should investigate individual differences in the perception of robots.

5.2 Ethical considerations

Our tools have the potential to uncover implicit biases that carry over from human-human interactions to human-robot interactions. For instance, our study revealed that participants tended to use tags like "playful" and "friendly" when describing a female robot and voice, whereas "assertive" and "functional" were more commonly associated with a male robot's image. Furthermore, when describing a male voice, participants frequently employed tags like "unemotional" and "reserved".

To assess if the correlational structure obtained here generalizes to other datasets, we repeated the dense rating experiment on new image and voice data and found strong correlations across the new and old datasets ($r = .85$ and $r = .91$ respectively). While this shows that our findings are robust across datasets, it does not rule out the possibility that both datasets are biased in the same way. To quantify the effect of the human-in-the-loop approach on perceived biases, we took, as an example, perceived gender. Consistent with the observation by Perugia and colleagues, we observe an underrepresentation of perceived female robots in both image datasets [60] (13 % for IEEE robots and 19 % for ABOT, percent below the midpoint for scale; see Supplementary Materials C.3).

However, when taking random voice samples from the text-to-speech model, participants evaluate the perceived gender of samples as nearly balanced (50 %). This is consistent with the dataset the model was trained on [92], which was intended to contain a diverse set of voices. Importantly, the percentage of perceived females in the GSP-matched voices was similarly balanced (49 %). This means that despite viewing an unbalanced dataset of images, the human-the-loop approach provided a much more balanced voice distribution. Future research can use more balanced sets of robot images or use GSP with a generative model to create images of robots, which would lead to the development of "customizable robots" as proposed by Schiebinger [71].

Next to potential dataset bias, we explored how implicit biases of participants actively can be reduced. Before rerunning the STEP and dense rating experiments, participants undergo implicit bias training and take the implicit association test (IAT) [22, 73]. While participants carefully read the implicit bias fact sheet (6/8 text-comprehension questions were answered correctly), we did not measure a significant decrease in bias. As a consequence, the intervention did not substantially alter the responses. This may be explained by the fact that the effects of the training are short-lived and hence barely change the implicit biases [41]. Future research can consider other interventions to reduce biases in the data. For example, a common theme in the development of recent Large Language Models is to perform an additional refinement on the model to suppress averse responses using human supervision [59]. We can perform an analog step for our approach, we can add a final post-processing step in which humans are asked to flag implicit biases.

Finally, we want to emphasize that the relationships across the terms we uncovered within and across two modalities are not causal, but are merely correlations. So, the fact that images of female-looking robots tend to be perceived as "cute" does not mean that they are cute because they are female (e.g., female-looking robots might have a more fluffy appearance on average, which makes them look cute).

5.3 General conclusion

The relationship between a robot's voice and its impact on user perception is complex. The primary aim of this paper was to explore the impact of nuanced voice features on individual perceptions of a diverse array of robots. In contrast to prior research – that used a small number of robots, existing audio samples, and limited behavioral testing – we adopted a multi-method approach that included generative AI, human-in-the-loop computations, and voice prediction. Our voice generation tool combines state-of-the-art TTS with traditional signal processing. We used human-in-the-loop computations in two key points in the research program: to collectively navigate a space of voice dimensions and to provide open-ended labeling of images and voices. We complement human-in-the-loop experiments with extensive behavioral validation experiments (N = 2,505). Our results demonstrated that participants consistently converged towards specific voice prototypes that either enhanced or aligned with the attributes associated with the static images of the robots. Our findings highlight the significant interplay between visual and auditory perceptions in shaping how humans perceive and attribute qualities to robotic entities. Furthermore, our study revealed that predicting a suitable voice for images of previously unseen robots is possible. This discovery can be interpreted as evidence that static visual cues alone may suffice to empower individuals to create voices that consistently convey the collective mental model of the respective robot. Using the perceptual dimensions we obtained for the set of robots, we could propose suitable voices to designers for new robots, as well as reveal and possibly suppress societal stereotypes underlying participants' choices. They are of practical relevance for engineers who want to fit a voice that matches a robot. More broadly, our research demonstrates the synergy between cognitive science and machine learning in tackling engineering challenges, such as human-robot interaction.

ACKNOWLEDGMENTS

REFERENCES

[1] 1969. *IEEE recommended practice for speech quality measurements.* Technical Report. IEEE. https://doi.org/10.1109/IEEESTD.1969.7405210 ISBN: 9781504402743.

[2] Fernando Alonso-Martín, María Malfaz, Álvaro Castro-González, José Carlos Castillo, and Miguel A Salichs. 2019. Online evaluation of text to speech systems for three social robots. In *International Conference on Social Robotics*. Springer, 155–164.

[3] Fernando Alonso Martin, María Malfaz, Álvaro Castro-González, José Carlos Castillo, and Miguel Ángel Salichs. 2020. Four-features evaluation of text to speech systems for three social robots. *Electronics* 9, 2 (2020), 267.

[4] Manuel Anglada-Tort, Peter M. C. Harrison, and Nori Jacoby. 2022. Studying the Effect of Oral Transmission on Melodic Structure using Online Iterated Singing Experiments. (May 2022). https://doi.org/10.1101/2022.05.10.491366

[5] Matthew P Aylett, Selina Jeanne Sutton, and Yolanda Vazquez-Alvarez. 2019. The right kind of unnatural: designing a robot voice. In *Proceedings of the 1st international conference on conversational user interfaces*. 1–2.

[6] Rainer Banse and Klaus R. Scherer. 1996. Acoustic Profiles in Vocal Emotion Expression. *Journal of Personality and Social Psychology* 70, 3 (1996), 614–636. https://doi.org/10/fnrx63

[7] H. Clark Barrett. 2020. Towards a Cognitive Science of the Human: Cross-Cultural Approaches and Their Urgency. *Trends in Cognitive Sciences* 24, 8 (Aug. 2020), 620–638. https://doi.org/10.1016/j.tics.2020.05.007

[8] 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 (Jan. 2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[9] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

[10] Paul Boersma and David Weenink. 2018. Praat: doing phonetics by computer [Computer program]. Version 6.0.37, http://www.praat.org/.

[11] Cynthia Breazeal. 2004. *Designing Sociable Robots.* MIT Press.

[12] Julia Cambre and Chinmay Kulkarni. 2019. One Voice Fits All?: Social Implications and Research Challenges of Designing Voices for Smart Devices. *Proc. ACM Hum. Comput. Interact.* 3, CSCW (2019), 223:1–223:19. https://doi.org/10.1145/3359325

[13] Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation. arXiv:2308.11596 [cs.CL]

[14] Marco Costa, Pio Enrico Ricci Bitti, and Luisa Bonfiglioli. 2000. Psychological connotations of harmonic musical intervals. *Psychology of Music* 28, 1 (2000), 4–22.

[15] Boele De Raad, Dick PH Barelds, Marieke E Timmerman, Kim De Roover, Boris Mlačić, and A Timothy Church. 2014. Towards a pan-cultural personality structure: Input from 11 psycholexical studies. *European Journal of Personality* 28, 5 (2014), 497–510.

[16] Ron Dotsch and Alexander Todorov. 2011. Reverse Correlating Social Face Perception. *Social Psychological and Personality Science* 3, 5 (dec 2011), 562–571. https://doi.org/10.1177/1948550611430272

[17] Anna Esposito, Terry Amorese, Marialucia Cuciniello, Maria Teresa Riviello, Antonietta Maria Esposito, Alda Troncone, and Gennaro Cordasco. 2019. The Dependability of Voice on Elders' Acceptance of Humanoid Agents.. In *Interspeech*. 31–35.

[18] Florian Eyben, Klaus R. Scherer, Bjorn W. Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 7, 2 (April 2016), 190–202. https://doi.org/10/f3sfxq

[19] Susan R. Fussell, Sara B. Kiesler, Leslie D. Setlock, and Victoria Yew. 2008. How people anthropomorphize robots. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction, HRI 2008, Amsterdam, The Netherlands, March 12-15, 2008*, Terry Fong, Kerstin Dautenhahn, Matthias Scheutz, and Yiannis Demiris (Eds.). ACM, 145–152. https://doi.org/10.1145/1349822.1349842

[20] Jennifer Goetz, Sara Kiesler, and Aaron Powers. 2003. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.* Ieee, 55–60.

[21] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality* 37, 6 (2003), 504 – 528. https://doi.org/10.1016/S0092-6566(03)00046-1

[22] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* 74, 6 (1998), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

[23] Markus Häring, Dieta Kuchenbrandt, and Elisabeth André. 2014. Would you like to play with me?: how robots' group membership and task features influence human-robot interaction. In *ACM/IEEE International Conference on Human-Robot Interaction, HRI'14, Bielefeld, Germany, March 3-6, 2014*, Gerhard Sagerer, Michita Imai, Tony Belpaeme, and Andrea Lockerd Thomaz (Eds.). ACM, 9–16. https://doi.org/10.1145/2559636.2559673

[24] Peter Harrison, Raja Marjieh, Federico Adolfi, Pol van Rijn, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, and Nori Jacoby. 2020. Gibbs sampling with people. *Advances in Neural Information Processing Systems* 33 (2020), 10659–10671.

[25] Marc Hassenzahl. 2004. The Interplay of Beauty, Goodness, and Usability in Interactive Products. *Human–Computer Interaction* 19, 4 (2004), 319–349. https://doi.org/10.1207/s15327051hci1904_2

[26] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & Computer 2003*, Gerd Szwillus and Jürgen Ziegler (Eds.). Vol. 57. Vieweg+Teubner Verlag, Wiesbaden, 187–196. https://doi.org/10.1007/978-3-322-80058-9_19

[27] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. Most people are not WEIRD. *Nature* 466, 7302 (June 2010), 29–29. https://doi.org/10.1038/466029a

[28] Willem K Hofstee, Boele De Raad, and Lewis R Goldberg. 1992. Integration of the big five and circumplex approaches to trait structure. *Journal of personality and social psychology* 63, 1 (1992), 146.

[29] Nori Jacoby and Josh H. McDermott. 2017. Integer Ratio Priors on Musical Rhythm Revealed Cross-culturally by Iterated Reproduction. *Current Biology* 27, 3 (2017), 359–370. https://doi.org/10.1016/j.cub.2016.12.031

[30] Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics* 71 (2018), 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

[31] Jesin James, BT Balamurali, Catherine I Watson, and Bruce MacDonald. 2020. Empathetic speech synthesis and testing for healthcare robots. *International Journal of Social Robotics* (2020), 1–19.

[32] Kathrin Janowski, Hannes Ritschel, and Elisabeth André. 2022. Adaptive Artificial Personalities. In *The Handbook on Socially Interactive Agents*. ACM, 155–194. https://doi.org/10.1145/3563659.3563666

[33] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems* 31 (2018).

[34] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems* 33 (2020), 8067–8077.

[35] Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*. PMLR, 5530–5540.

[36] Jaehyeon Kim, Jungil Kong, and Juhee Son. 2022. VITS implementation. https://github.com/jaywalnut310/vits.

[37] Klaus J. Kohler. 2011. Communicative Functions Integrate Segments in Prosodies and Prosodies in Segments. *Phonetica* 68, 1-2 (2011), 26–56. https://doi.org/10/d9vhxq

[38] Daichi Kondo and Masanori Morise. 2019. Human-in-the-loop speech-design system and its evaluation. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 608–612.

[39] Dieta Kuchenbrandt, Markus Häring, Jessica Eichberg, Friederike Eyssel, and Elisabeth André. 2014. Keep an Eye on the Task! How Gender Typicality of Tasks Influence Human-Robot Interactions. *Int. J. Soc. Robotics* 6, 3 (2014), 417–427. https://doi.org/10.1007/s12369-014-0244-0

[40] Katharina Kühne, Martin H Fischer, and Yuefang Zhou. 2020. The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study. *Frontiers in neurorobotics* (2020), 105.

[41] Calvin K. Lai, Maddalena Marini, Steven A. Lehr, Carlo Cerruti, Jiyun-Elizabeth L. Shin, Jennifer A. Joy-Gaba, Arnold K. Ho, Bethany A. Teachman, Sean P. Wojcik, Spassena P. Koleva, Rebecca S. Frazier, Larisa Heiphetz, Eva E. Chen, Rhiannon N. Turner, Jonathan Haidt, Selin Kesebir, Carlee Beth Hawkins, Hillary S. Schaefer, Sandro Rubichi, Giuseppe Sartori, Christopher M. Dial, N. Sriram, Mahzarin R. Banaji, and Brian A. Nosek. 2014. Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General* 143, 4 (2014), 1765–1785. https://doi.org/10.1037/a0036260

[42] Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6588–6592.

[43] Thomas Langlois, Nori Jacoby, Jordan W. Suchow, and Tom Griffiths. 2019. Orthogonal multi-view three-dimensional object representations in memory revealed by serial reproduction. In *Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation, Montreal, Canada, July 24-27, 2019*, Ashok K. Goel, Colleen M. Seifert, and Christian Freksa (Eds.). cognitivesciencesociety.org, 2078–2083. https://mindmodeling.org/cogsci2019/papers/0363/index.html

[44] Thomas A. Langlois, Nori Jacoby, Jordan W. Suchow, and Thomas L. Griffiths. 2021. Serial reproduction reveals the geometry of visuospatial representations. *Proceedings of the National Academy of Sciences* 118, 13 (March 2021). https://doi.org/10.1073/pnas.2012938118

[45] Michael C. Mangini and Irving Biederman. 2004. Making the ineffable explicit: estimating the information employed for face classifications. *Cognitive Science* 28, 2 (mar 2004), 209–226. https://doi.org/10.1207/s15516709cog2802_4

[46] Raja Marjieh, Pol van Rijn, Ilia Sucholutsky, Theodore R. Sumers, Harin Lee, Thomas L. Griffiths, and Nori Jacoby. 2022. Words are all you need? Capturing human sensory similarity with textual descriptors. *arXiv preprint arXiv:2206.04105* (2022).

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

[48] RR McCrae and OP John. 1992. An introduction to the five-factor model and its applications. *Journal of personality* 60, 2 (1992), 175.

[49] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, Vol. 8. Citeseer, 18–25.

[50] Conor McGinn and Ilaria Torre. 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, 211–221.

[51] Lianne F. S. Meah and Roger K. Moore. 2014. *The Uncanny Valley: A Focus on Misaligned Cues*. Springer International Publishing, 256–265. https://doi.org/10.1007/978-3-319-11973-1_26

[52] Silvan Mertes, Thomas Kiderle, Ruben Schlagowski, Florian Lingenfelser, and Elisabeth Andre. 2021. On the potential of modular voice conversion for virtual agents. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 1–7.

[53] Alice E Milne, Roberta Bianco, Katarina C Poole, Sijia Zhao, Andrew J Oxenham, Alexander J Billig, and Maria Chait. 2021. An online headphone screening test based on dichotic pitch. *Behavior Research Methods* 53, 4 (2021), 1551–1562.

[54] Wade J Mitchell, Kevin A Szerszen, Amy Shirong Lu, Paul W Schermerhorn, Matthias Scheutz, and Karl F MacDorman. 2011. A Mismatch in the Human Realism of Face and Voice Produces an Uncanny Valley. *i-Perception* 2, 1 (Jan. 2011), 10–12. https://doi.org/10.1068/i0415

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

[56] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review* 56, 4 (2023), 3005–3054.

[57] Bilge Mutlu, Steven Osman, Jodi Forlizzi, Jessica K. Hodgins, and Sara B. Kiesler. 2006. Task Structure and User Attributes as Elements of Human-Robot Interaction Design. In *The 15th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2006, Hatfield, Herthfordshire, UK, 6-8 September, 2006*. IEEE, 74–79. https://doi.org/10.1109/ROMAN.2006.314397

[58] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Conference on Human Factors in Computing Systems, CHI 1994, Boston, Massachusetts, USA, April 24-28, 1994, Proceedings*, Beth Adelson, Susan T. Dumais, and Judith S. Olson (Eds.). ACM, 72–78. https://doi.org/10.1145/191666.191703

[59] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[60] Giulia Perugia, Stefano Guidi, Margherita Bicchi, and Oronzo Parlangeli. 2022. The Shape of Our Bias: Perceived Age and Gender in the Humanoid Robots of the ABOT Database. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. https://doi.org/10.1109/hri53351.2022.9889366

[61] Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F. Malle. 2018. What is Human-like?: Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. ACM. https://doi.org/10.1145/3171221.3171268

[62] Emmanuel Ponsot, Juan José Burred, Pascal Belin, and Jean-Julien Aucouturier. 2018. Cracking the Social Code of Speech Prosody Using Reverse Correlation. *Proceedings of the National Academy of Sciences* 115, 15 (March 2018), 3972–3977. https://doi.org/10.1073/pnas.1716090115

[63] Aaron Powers and Sara Kiesler. 2006. The advisor robot: tracing people's mental model from a robot's physical attributes. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 218–225.

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[65] Nicolas Ramirez-Guevara. 2017. Robotization Effect Using Phase Vocoder Processing. (2017).

[66] Beatrice Rammstedt and Oliver P. John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality* 41, 1 (2007), 203 – 212. https://doi.org/10.1016/j.jrp.2006.02.001

[67] Hannes Ritschel, Ilhan Aslan, Silvan Mertes, Andreas Seiderer, and Elisabeth André. 2019. Personalized synthesis of intentional and emotional non-verbal sounds for social robots. In *2019 8th International conference on affective computing and intelligent interaction (ACII)*. IEEE, 1–7.

[68] Sigrid Roehling, Bruce MacDonald, and Catherine Watson. 2006. Towards expressive speech synthesis in english on a robotic platform. In *Proceedings of the Australasian International Conference on Speech Science and Technology*. Citeseer, 130–135.

[69] Adam N. Sanborn, Thomas L. Griffiths, and Richard M. Shiffrin. 2010. Uncovering mental representations with Markov chain Monte Carlo. *Cognitive Psychology* 60, 2 (mar 2010), 63–106. https://doi.org/10.1016/j.cogpsych.2009.07.001

[70] Klaus R. Scherer. 1978. Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology* 8, 4 (Oct. 1978), 467–487. https://doi.org/10.1002/ejsp.2420080405

[71] Londa Schiebinger. 2019. The robots are coming! But should they be gendered. *AWIS Magazine, Winter* (2019), 18–21, 58.

[72] Dominik Schiller, Silvan Mertes, Pol van Rijn, and Elisabeth André. 2021. Analysis by Synthesis: Using an Expressive TTS Model as Feature Extractor for Paralinguistic Speech Classification. In *Proc. Interspeech 2021*. 486–490. https://doi.org/10.21437/Interspeech.2021-1587

[73] Konrad Schnabel, Jens B. Asendorpf, and Anthony G. Greenwald. 2008. *Using Implicit Association Tests for the Assessment of Implicit Personality Self-Concept*. SAGE Publications Ltd, 508–528. https://doi.org/10.4135/9781849200479.n24

[74] Valentin Schwind. 2018. Implications of the uncanny valley of avatars and virtual characters for human-computer interaction. https://doi.org/10.18419/OPUS-9936

[75] Hubert Siuzdak, Piotr Dura, Pol van Rijn, and Nori Jacoby. 2022. WavThruVec: Latent speech representation as intermediate features for neural speech synthesis. https://doi.org/10.48550/ARXIV.2203.16930

[76] Daisy Stanton, Matt Shannon, Soroosh Mariooryad, RJ Skerry-Ryan, Eric Battenberg, Tom Bagby, and David Kao. 2021. Speaker Generation. *arXiv preprint arXiv:2111.05095* (2021).

[77] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A Survey on Neural Speech Synthesis. https://doi.org/10.48550/ARXIV.2106.15561

[78] Angela Tinwell, Mark Grimshaw, and Deborah Abdel Nabi. 2015. The effect of onset asynchrony in audio-visual speech and the Uncanny Valley in virtual

characters. *International Journal of Mechanisms and Robotic Systems* 2, 2 (2015), 97. https://doi.org/10.1504/ijmrs.2015.068991

[79] Ilaria Torre, Jeremy Goslin, and Laurence White. 2020. If your device could smile: People trust happy-sounding artificial agents more. *Computers in Human Behavior* 105 (2020), 106215.

[80] Ilaria Torre, Jeremy Goslin, Laurence White, and Debora Zanatto. 2018. Trust in artificial voices: A" congruency effect" of first impressions and behavioural experience. In *Proceedings of the Technology, Mind, and Society*. 1–6.

[81] Paul D. Trapnell and Jerry S. Wiggins. 1990. Extension of the Interpersonal Adjective Scales to include the Big Five dimensions of personality. *Journal of Personality and Social Psychology* 59, 4 (1990), 781–790. https://doi.org/10.1037/0022-3514.59.4.781

[82] Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro. 2020. Mellotron: Multispeaker Expressive Voice Synthesis by Conditioning on Rhythm, Pitch and Global Style Tokens. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6189–6193. https://doi.org/10/gg3jcz

[83] Rafael Valle, Kevin Shih, Ryan Prenger, and Bryan Catanzaro. 2020. Flowtron: An Autoregressive Flow-Based Generative Network for Text-to-Speech Synthesis. *arXiv:2005.05957 [cs, eess]* (2020). arXiv:2005.05957 [cs, eess]

[84] Pol van Rijn, Harin Lee, and Nori Jacoby. 2022. Bridging the prosody GAP: Genetic Algorithm with People to efficiently sample emotional prosody, In CogSci. *arXiv preprint arXiv:2205.04820*.

[85] Pol van Rijn, Silvan Mertes, Dominik Schiller, Piotr Dura, Hubert Siuzdak, Peter M. C. Harrison, Elisabeth André, and Nori Jacoby. 2022. VoiceMe: Personalized voice generation in TTS. In *Proc. Interspeech 2022*. 2588–2592. https://doi.org/10.21437/Interspeech.2022-10855

[86] Pol van Rijn, Silvan Mertes, Dominik Schiller, Peter M.C. Harrison, Pauline Larrouy-Maestri, Elisabeth André, and Nori Jacoby. 2021. Exploring Emotional Prototypes in a High Dimensional TTS Latent Space. In *Proc. Interspeech 2021*. 3870–3874. https://doi.org/10.21437/Interspeech.2021-1538

[87] Pol van Rijn, Yue Sun, Harin Lee, Raja Marjieh, Ilia Sucholutsky, Francesca Lanzarini, Elisabeth André, and Nori Jacoby. 2023. Around the world in 60 words: A generative vocabulary test for online research. arXiv:arXiv:2302.01614

[88] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Verena Winterhalter, Markus Bühner, and Heinrich Hussmann. 2020. Developing a Personality Model for Speech-Based Conversational Agents Using the Psycholexical Approach. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI ’20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376210 event-place: Honolulu, HI, USA.

[89] Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous. 2018. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. *arXiv:1803.09017 [cs, eess]* (2018). arXiv:1803.09017 [cs, eess]

[90] Kevin JP Woods, Max H Siegel, James Traer, and Josh H McDermott. 2017. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics* 79, 7 (2017), 2064–2072.

[91] Jing Xu, Mike Dowman, and Thomas L. Griffiths. 2013. Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences* 280, 1758 (May 2013), 20123073. https://doi.org/10.1098/rspb.2012.3073

[92] Junichi Yamagishi, Christophe Veaux, and Kirsten. MacDonald. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). https://doi.org/10.7488/DS/2645

A CODE AND DATA AVAILABILITY

A view-only anonymous link is provided to the public, containing all the data collected for this project during the review stage [6]. It includes the new human behavioral data, the computational experiments with machine learning models, and all the necessary analysis scripts for producing the results. Additionally, the repository includes the PsyNet source codes for reproducing the behavioral experiments. Finally, we present an interactive visualization [7] for exploring the created voices, the perceptual space of robots and for predicting new voices based on the perceptual dimensions.

B BEHAVIORAL STUDIES

B.1 Participants

Participants were recruited from Prolific[8] and provided informed consent under an approved protocol. The median age was 38 (SD: 12.6, min: 18, max: 88). 61.4 % of the participants identified themselves as male, 36.6 as female, 0.1 % as non-binary and 0.1 % preferred not to say. The highest level of formal education is high school for 23.3 %, college for 34.8 %, graduate school for 25.1 %, and postgraduate school or higher for 16.7 % of the participants (0.1 % of the participants had no formal education). The exact number of participants for each of the 7 behavioral experiments is reported in Table S1.

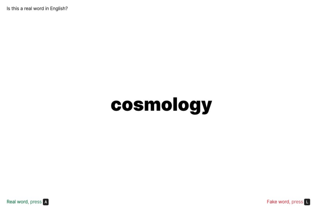
The median total durations of the experiments are typical for online experiments. The duration of the GSP experiments are similar to other experiments using GSP [S24, S85, S86].

B.2 Implementation

All behavioral experiments were implemented using PsyNet framework [S24]. PsyNet is a novel experiment design framework that builds on Dallinger (https://dallinger.readthedocs.io/) and allows for flexible specification of experiment timelines as well as providing support for a wide array of tasks across different modalities (visual, auditory, and audio-visual). Dallinger is a modern tool for experiment hosting and deployment that automates the process of participant recruitment and compensation by integrating cloud-based services such as Heroku[9] with online crowd-sourcing platforms such as Prolific. Participants interact with the experiment through their web browser, which in turn communicates with a backend Python server responsible for the experiment logic. As an advantage of using PsyNet, it offers native support for adaptive human-in-the-loop experiments.

B.3 Pre-screening

In order to collect high-quality data, pre-screening tasks were used to avoid low-quality participants and users who used bots to respond. We conduct the pre-screeners right before the main experiment. If the pre-screening tasks are not completed, the experiment will be terminated early, but the participants will still be paid for their time (regardless of the outcome). Pre-screeners are additionally ensure two main criteria for data quality, namely, a) to ensure that participants are wearing headphones and can hear audio b) that they are native speakers of the language. To do this, we implemented two tasks from previous literature. Namely, an English proficiency test ([S87]) for experiments that relied on text; and a standardized headphone test ([S90] used for experiments involving audio. Table S1 provides details on which pre-screeners were used in each of the behavioral experiments.

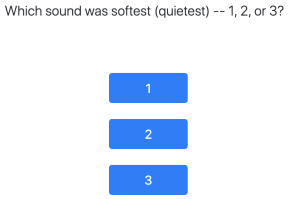


Is this a real word in English?

cosmology

Figure S1: Example trial from the WikiVocab pre-screening task [S87].

English proficiency test. To test participants' English proficiency we used the lexical decision task WikiVocab [S87]. In each trial, we briefly present the participant (1 second) with either a real English word or a pseudo-word that does not exist. Participants were instructed to guess whether the word was real or not. They used dedicated keys on their keyboard to respond. A total of 30 trials (half of them being real words) were presented, and 25 of them needed to be correct for the participant to pass. For each batch of 30 trials, we randomly selected 15 real and 15 fake words. See https://vocabtest.org/ for an implementation of the test. An example trial is shown in Figure S1.



Which sound was softest (quietest) -- 1, 2, or 3?

Figure S2: Example trial from the headphone pre-screening test [S90].

Headphone test. We used the headphone test developed by Wood et al. [S90], which is used as a standard pre-screener for high-quality auditory psychophysics data-collection procedures [S53]. The test is designed to ensure that the participants are wearing headphones and are able to perceive subtle differences in volume. The task consists of a forced choice task, in which three consecutive tones are played, and the participant has to identify which of them is the quietest. Importantly, these tones exhibit a phase cancellation effect without headphones, making it difficult for non-headphone users to identify the quietest tone. To pass, participants had to answer 4 out of 6 trials correctly. An example trial is shown in Figure S2.

[6] **Code and data:** https://robotvoice.s3.amazonaws.com/supplementary_materials.zip
[7] **Interactive plots:** https://robotvoice.s3.amazonaws.com/index.html
[8] https://www.prolific.co/
[9] https://www.heroku.com/

Table S1: Behavioral experiment summary table. *Note.* *N* denotes the number of participants included in the analysis; WV denotes the WikiVocab English proficiency pre-screening task [S87]; HT denotes the headphone test [S90]. * means that before the main experiment, participants did implicit bias awareness training. Type: M denotes an experiment for the main results, C denotes a control experiment. Dur. denotes the median duration in minutes.

Modality	Paradigm	Total stimuli	Trials per participant	Section	*N*	Pre-screening	Type	Dur.
Voices + Images	GSP	175	20	4.1	803	HT	M	11.8
Voices + Images	Validation	3,255	80	4.1	142	HT	M	15.3
Images	STEP-Tag	175	30	4.2	73	WV	M	12.7
Images	STEP-Tag*	175	30	4.2	78	WV	C	19.9
Voices	STEP-Tag	175	30	4.2	59	HT, WV	M	12.0
Images	Rating	175	60	4.3	298	WV	M	13.5
Images*	Rating	175	60	4.3	202	WV	C	20.8
Voices	Rating	175	60	4.3	245	HT, WV	M	14.1
New Images	Rating	175	60	4.3	249	WV	C	14.5
Random Voices	Rating	175	60	4.3	189	HT, WV	C	15.0
Voices + Images	Prediction	875	80	4.4	94	HT	M	10.1
Voices + New images	Prediction new	700	80	4.4	73	HT	C	11.2

C METHODS

C.1 Selection of images

We selected a wide variety of different robots from the robot database IEEE Robots. The images span 14 categories as marked by the database (see Figure S3). The majority of the images fall into the categories "humanoid" and "research".

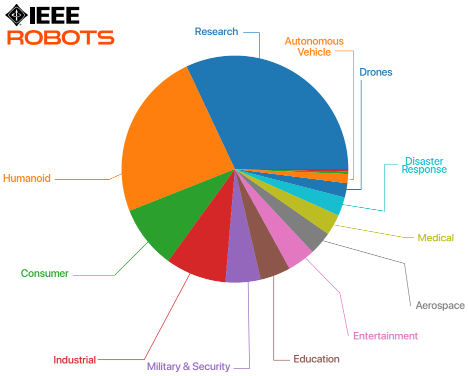


Figure S3: Distribution of different categories of robots used from IEEE Robots.

Tables S9–S10 shows examples of the images selected and edited for the experiment. The full list of stimuli is available at https://s3.amazonaws.com/robotvoice/explore.html.

C.2 Selection of new set of images

To assess the generalizability of our findings, we ran the dense rating on a new set of images and voices (see section F.5). We used the ABOT robot database[10] [S60, S61] and obtained 167 new robot images from it after removing those that were already in the IEEE Robots selection.

We then added the following 7 robots that were used in a study by Mathur and Reichling [S47] but not already included in the other two lists.

- 3e A18 (Honda)[11]
- 3e C18 (Honda)[12]
- aeo (Aeolus Robotics)[13]
- cruzr (Ubtech Robotics)[14]
- Jules (Hanson Robotics)[15]
- Actroid Repliee Q2 (Osaka University, Kokoro Co. Ltd)[16]
- Tapia (MJI Robotics)[17]

Another robot that we added was Emotech's Olly [18] because its abstract design is very different from that of the other robots and seemingly contrasts with the personality that its creators emphasized in advertising.

C.3 Selection bias in images

Consistent with previous literature, we find an underrepresentation of female robots in the ABOT database [S60] and in the IEEE Robot database (see Figure S4A). The initial random robot voices are gender-balanced (Figure S4B). Interestingly, the matched voices to the biased sample of robot images are equally gender-balanced as the initial random voices.

C.4 Latent voice dimensions

In an initial set of pilots, we noted that using only the TTS system to create voices provides diverse human-like voices but does not allow

[10] https://www.abotdatabase.info/

[11] https://global.honda/en/innovation/CES/2018/001.html

[12] https://www.honda.com.au/news/2018/honda-3e-robot-concept

[13] https://aeolusbot.com/

[14] https://ubtrobot.com/

[15] https://www.hansonrobotics.com/jules/

[16] https://en.wikipedia.org/wiki/Actroid

[17] http://mjirobotics.co.jp/en

[18] https://www.indiegogo.com/projects/olly-the-first-home-robot-with-personality/

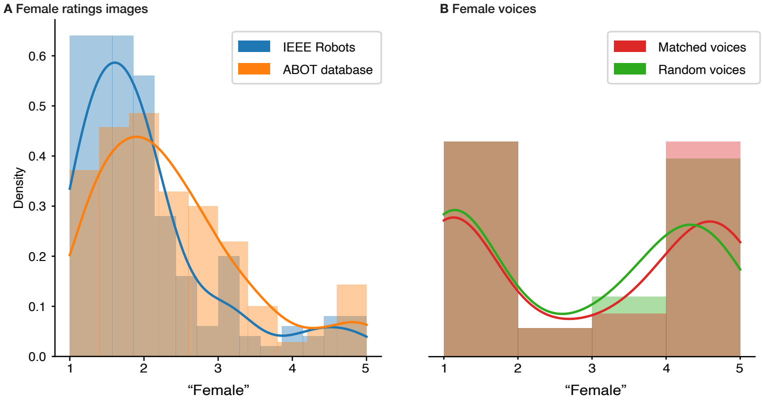


Figure S4: Distribution of ratings for attribute "female" in image (A) and voice datasets (B).

us a way to create mechanical voices. Thus, adding effects was essential. Once we add effects, these effects can also account for some of the voice characteristics (e.g., pitch height, speech duration, voice roughness, etc.). This decreases the importance of having very high dimensional TTS representations. We experimented with various dimension reduction algorithms. In particular, we experimented with supervised (PLS and CCA) and unsupervised dimension techniques (PCA). For PLS and CCA, we used the eGeMAPS feature set [S18] as a supervision signal. Based on initial piloting, we found that the PCA voice dimensions had the best tradeoff between expressivity and distortion (high expressivity, low distortion). Across the dimension reduction methods, we found that the total amount of explained variance was comparable from method to method and was generally relatively small. This is compatible with previous studies in which the explained variance of the PCA on latent voice dimensions was low [S85] (e.g., 25.4 % variance explained for ten dimensions). We reduced the total number of dimensions to five dimensions, capturing 12.2 % of the variance (see Figure S5). Note that the explained variance per dimension is high for the first few dimensions and decays slowly afterward, which makes the choice of five dimensions reasonable. Moreover, our pilot suggested that adding further dimensions did not improve voice expressivity qualitatively. Since most of the sliders in the experiment are voice dimensions (5/8), using a lower number of voice dimensions allows us to revisit dimensions more often in the GSP process and accelerate convergence. While we only use five dimensions, the dimensions clearly capture various aspects of the voice, such as sex and age (see Supplementary Materials F.6).

C.5 Voice Effects

We implemented a set of audio effects to allow the creation of voices that sound more synthetic. For each effect, the slider in our GSP experiment steered the amount of effect in the resulting signal. The only exception here is the *Timeshift* effect, where the slider did not control the amount of the effect but the time that the signal was shifted. In order to even out differences in auditory saliency between effects, we define upper bounds for the effect amount separately for each effect (see Table S2). These bounds were manually adjusted by one of the authors and further tested by all other authors. Further, some effects had to be parametrized with additional parameters which are listed in Table S3.

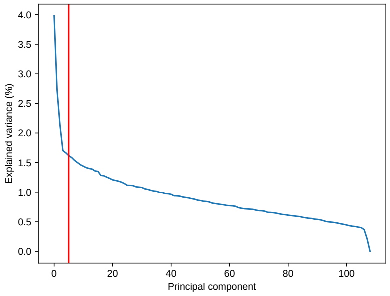


Figure S5: Explained variance by Principal Component Analysis.

C.6 Robot labels

The list of 260 robot labels is depicted in Table S4–S7. These labels were taken from psychological literature about human personality traits [S15, S21, S28, S48, S66, S81], from the Godspeed Questionnaires that were created specifically for social robots [S8], from the personality dimensions that Völkel et al. identified for voice assistants [S88], as well as from the AttrakDiff questionnaire that measures user experience [S25, S26]. The references for each label are listed next to it in the aforementioned tables.

In many cases, the sources did not contain the exact word, but a synonym or an antonym of it. These instances are marked with a ' respectively * symbol next to the source's abbreviation.

Furthermore, several labels were added after a small-scale pilot study for the labeling task showed a tendency for labeling robots with their visible properties. In particular, labels were added for the (apparent) sex, gender presentation and age since they were also expected to relate to the voices. Other examples include "animallike" due to the large number of non-human robots, as well as adjectives referring to the size or to attractiveness in general. In the tables, these are marked with a ○ symbol.

Effect	Upper Bound
Pitch	0.5
Tremolo	0.4
Synthesis Quality	1.0
Timeshift	45ms
Vocoder	0.35
Flanger	0.78

Table S2: Upper boundaries for the effect amounts for each effect. Lower boundaries is always 0.

Effect	Parameter	Value
Flanger Type 1	Delay	1
Flanger Type 1	Depth	10
Flanger Type 1	Frequency	5
Flanger Type 2	Delay	0
Flanger Type 2	Depth	50
Flanger Type 2	Frequency	0
Flanger Type 3	Delay	20
Flanger Type 3	Depth	20
Flanger Type 3	Frequency	5
Flanger Type 4	Delay	1
Flanger Type 4	Depth	10
Flanger Type 4	Frequency	25
Flanger Type 5	Delay	10
Flanger Type 5	Depth	0
Flanger Type 5	Frequency	0
Vocoder	Carrier Frequency	30
Vocoder	Harmonics	1.0

Table S3: Fixed parameters of the effects.

C.7 Number of robots in previous studies

The present study incorporates a much larger number of robots (350) than previous research (max. 8 different robots):

- Alonso-Martín et al. [S2]: 3 robots
- Alonso-Martín et al. [S3]: 3 robots
- Aylett et al. [S5]: 3 robots
- Häring et al. [S23]: 1 robot
- James et al. [S31]: 1 robot
- Kuchenbrandt et al. [S39]: 1 robot
- McGinn & Torre [S50]: 8 robots
- Powers & Kiesler [S63]: 4 robots
- Ritschel et al. [S67]: 1 robot
- Torre et al. [S80]: 1 robot

C.8 Implicit bias awareness training

We adapted the Implicit Association Test (IAT) [S73] to assess implicit biases in our participants. This test measures the association between words (either passive and active attributes or positive and negative words). The task is done on randomly selected six images from each target in a pair. Possible target pairs are adult vs. children, cats vs. dogs, and men vs. women. We focus on the target pair men vs. women, since the implicit biases were pronounced in the collected data. Since some attributes in the test are rarely used – such as "servile" or "obsequious" –, we select the 10 most frequent words from the 16 active and 16 passive words. The selected words occur at least once per one million. The selected active words are: "strong", "active", "effective", "mobile", "alive", "dynamic", "animated", "lively", "potent", and "energetic". The selected passive words are: "gentle", "passive", "inactive", "tame", "compliant", "yielding", "meek", "submissive", "obedient", and "controllable".

For each participant, we randomly select three passive and three active words. We tell the participant "In this task, you will be shown a word and two images. Your task is to choose the image that fits the word better." On every page, participants see a random male or female image (from the 10 images per gender in the IAT). The order of the female image (left or right) was random. On top of the image, participants see one of the 6 selected words. The images are shown for 2 seconds and automatically disappear. Participants use the keys on their keyboard to indicate if the left (key A) or the right image (key L) fits best to the attribute. Each of the six attributes is visited 5 times. After the participants complete all 30 trials, we show the measured biases (see Figure S6 for an example).

Participants proceed with an implicit bias training. For this, we selected the following excerpts from the implicit bias fact sheet from the White House Office of Science and Technology Policy[19] (Figure S7–S8)

In both experiments, on average, participants answered 6/8 questions correctly, indicating they carefully read the implicit bias awareness sheet. In both experiments, we did not find a significant difference across the terms after correcting for multiple comparisons (Bonferroni).

D CREATE VOICES USING GIBBS SAMPLING WITH PEOPLE

D.1 Instructions Main Experiment

The experiments proceeded as follows: Upon completion of the consent form and the pre-screening tasks, participants received instructions regarding the main experiment (see Figure S9).

As described in the instructions, in each trial, participants move a slider corresponding to one voice dimension and have to move the slider to the position such that the obtained voice maximally matches the robot. A screenshot of the task is shown in Figure S10.

[19] https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/bias_9-14-15_final.pdf

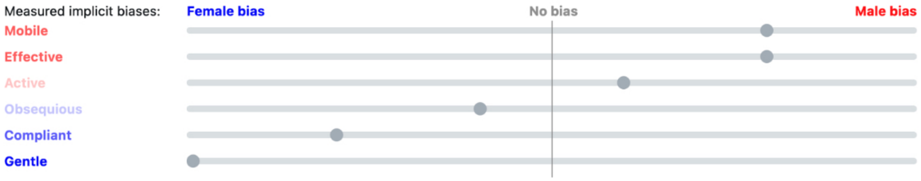
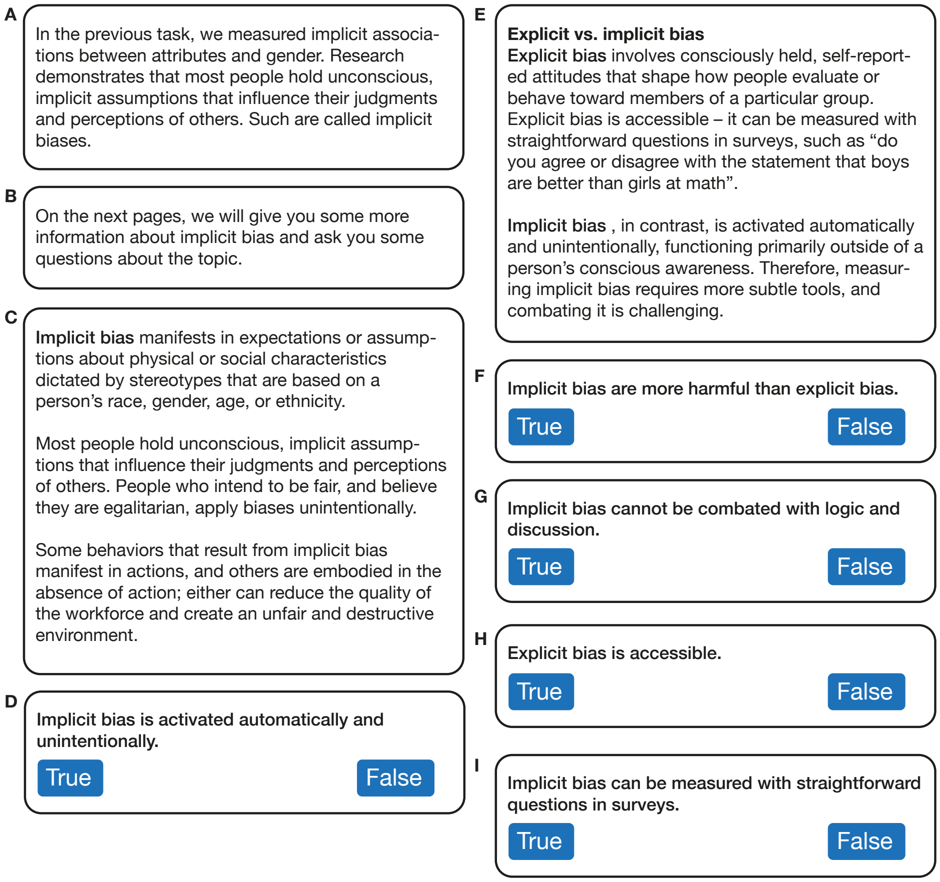


Figure S6: Example of measured biases in Implicit Association Test.



A

In the previous task, we measured implicit associations between attributes and gender. Research demonstrates that most people hold unconscious, implicit assumptions that influence their judgments and perceptions of others. Such are called implicit biases.

B

On the next pages, we will give you some more information about implicit bias and ask you some questions about the topic.

C

Implicit bias manifests in expectations or assumptions about physical or social characteristics dictated by stereotypes that are based on a person's race, gender, age, or ethnicity.

Most people hold unconscious, implicit assumptions that influence their judgments and perceptions of others. People who intend to be fair, and believe they are egalitarian, apply biases unintentionally.

Some behaviors that result from implicit bias manifest in actions, and others are embodied in the absence of action; either can reduce the quality of the workforce and create an unfair and destructive environment.

D

Implicit bias is activated automatically and unintentionally.

True False

E

Explicit vs. implicit bias
Explicit bias involves consciously held, self-reported attitudes that shape how people evaluate or behave toward members of a particular group. Explicit bias is accessible – it can be measured with straightforward questions in surveys, such as "do you agree or disagree with the statement that boys are better than girls at math".

Implicit bias , in contrast, is activated automatically and unintentionally, functioning primarily outside of a person's conscious awareness. Therefore, measuring implicit bias requires more subtle tools, and combating it is challenging.

F

Implicit bias are more harmful than explicit bias.

True False

G

Implicit bias cannot be combated with logic and discussion.

True False

H

Explicit bias is accessible.

True False

I

Implicit bias can be measured with straightforward questions in surveys.

True False

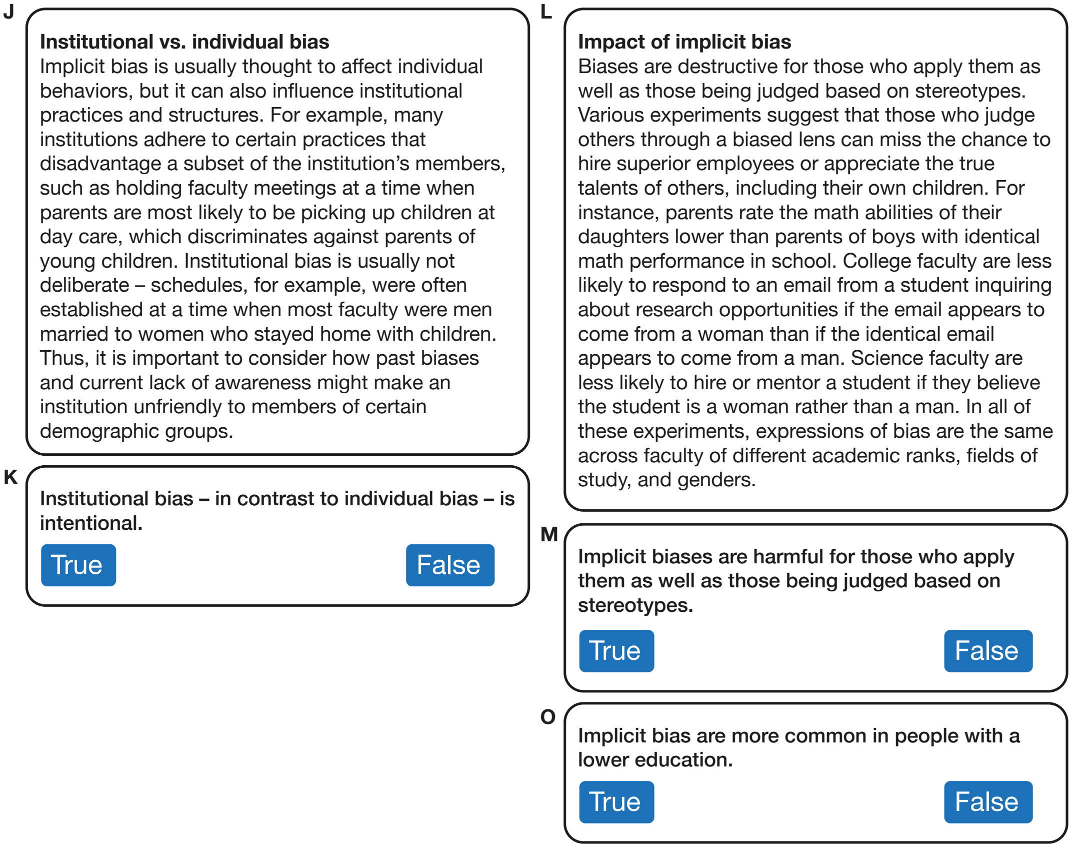
Figure S7: Implicit bias pages part 1. The letters indicate page order.

D.2 Different ways to summarize data from previous iterations in GSP experiments

Harrison et al. [S24] found that aggregating data from multiple participants in Gibbs Sampling with People (GSP) can improve sampling quality by reducing noise. However, the choice of how to summarize the aggregated data (e.g., mean, median, kernel density estimate) can impact the results. They used mean aggregation for a GSP experiment involving color-matching tasks (Experiment 1 in their paper) but selected the most common value, for face generation experiments (Experiment 4 in their paper), which they deemed more suitable for complex, multi-modal data. While we couldn't directly use their most common value aggregation approach due to a limited number of responses, we used a similar approach of median aggregation, which ensures that only played voice configurations propagate to the next iteration. Median aggregation, like the choice of the most frequent value in Harrison et al.'s generative face domain, is appropriate for our domain, because it prevents the selection of an unpopular intermediate value, as stimulus generation is time-consuming and slider changes may not be smooth.

D.3 Instructions Validation Experiment

The instructions for the experiment are shown in Figure S11. A screenshot of the task is shown in Figure S12.



Institutional vs. individual bias
Implicit bias is usually thought to affect individual behaviors, but it can also influence institutional practices and structures. For example, many institutions adhere to certain practices that disadvantage a subset of the institution's members, such as holding faculty meetings at a time when parents are most likely to be picking up children at day care, which discriminates against parents of young children. Institutional bias is usually not deliberate – schedules, for example, were often established at a time when most faculty were men married to women who stayed home with children. Thus, it is important to consider how past biases and current lack of awareness might make an institution unfriendly to members of certain demographic groups.

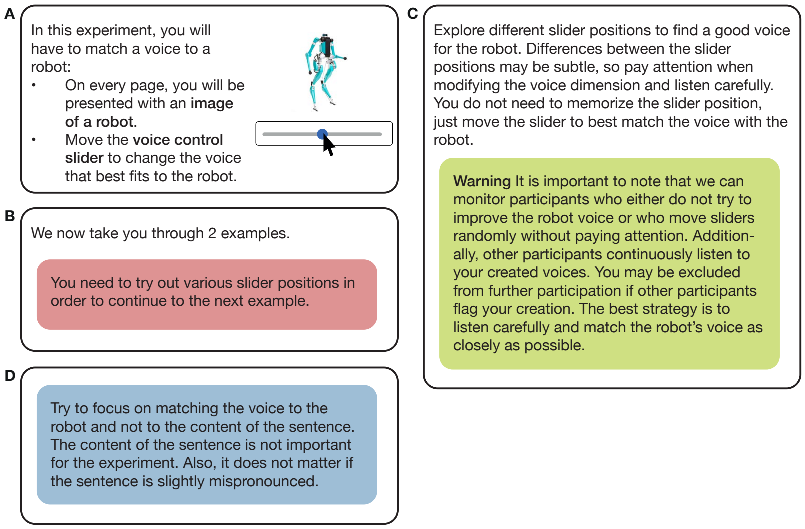
Institutional bias – in contrast to individual bias – is intentional.

Impact of implicit bias
Biases are destructive for those who apply them as well as those being judged based on stereotypes. Various experiments suggest that those who judge others through a biased lens can miss the chance to hire superior employees or appreciate the true talents of others, including their own children. For instance, parents rate the math abilities of their daughters lower than parents of boys with identical math performance in school. College faculty are less likely to respond to an email from a student inquiring about research opportunities if the email appears to come from a woman than if the identical email appears to come from a man. Science faculty are less likely to hire or mentor a student if they believe the student is a woman rather than a man. In all of these experiments, expressions of bias are the same across faculty of different academic ranks, fields of study, and genders.

Implicit biases are harmful for those who apply them as well as those being judged based on stereotypes.

Implicit bias are more common in people with a lower education.

Figure S8: Implicit bias pages part 2. The letters indicate page order.



In this experiment, you will have to match a voice to a robot:

- On every page, you will be presented with an **image of a robot**.
- Move the **voice control slider** to change the voice that best fits to the robot.

We now take you through 2 examples.

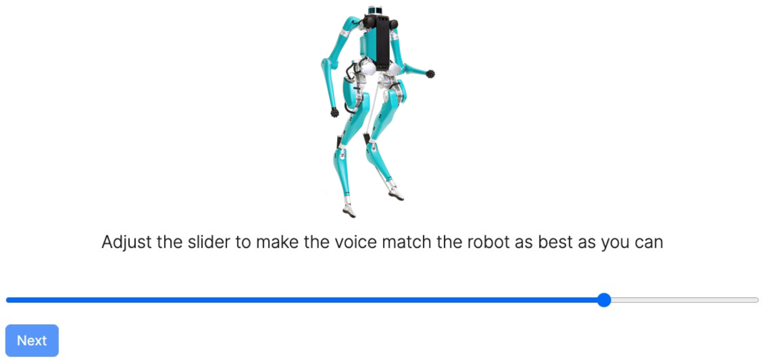
You need to try out various slider positions in order to continue to the next example.

Try to focus on matching the voice to the robot and not to the content of the sentence. The content of the sentence is not important for the experiment. Also, it does not matter if the sentence is slightly mispronounced.

Explore different slider positions to find a good voice for the robot. Differences between the slider positions may be subtle, so pay attention when modifying the voice dimension and listen carefully. You do not need to memorize the slider position, just move the slider to best match the voice with the robot.

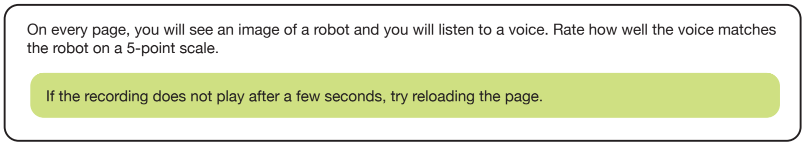
Warning It is important to note that we can monitor participants who either do not try to improve the robot voice or who move sliders randomly without paying attention. Additionally, other participants continuously listen to your created voices. You may be excluded from further participation if other participants flag your creation. The best strategy is to listen carefully and match the robot's voice as closely as possible.

Figure S9: Instructions for GSP experiment.



Adjust the slider to make the voice match the robot as best as you can

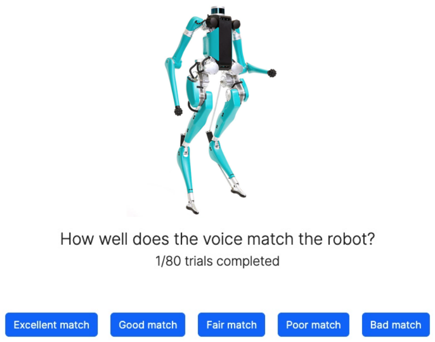
Figure S10: Example trial in the GSP experiment.



On every page, you will see an image of a robot and you will listen to a voice. Rate how well the voice matches the robot on a 5-point scale.

If the recording does not play after a few seconds, try reloading the page.

Figure S11: Instructions for validation experiment.



How well does the voice match the robot?

1/80 trials completed

Figure S12: Example trial in the validation experiment.

E ANNOTATE DIMENSION USING STEP-TAG

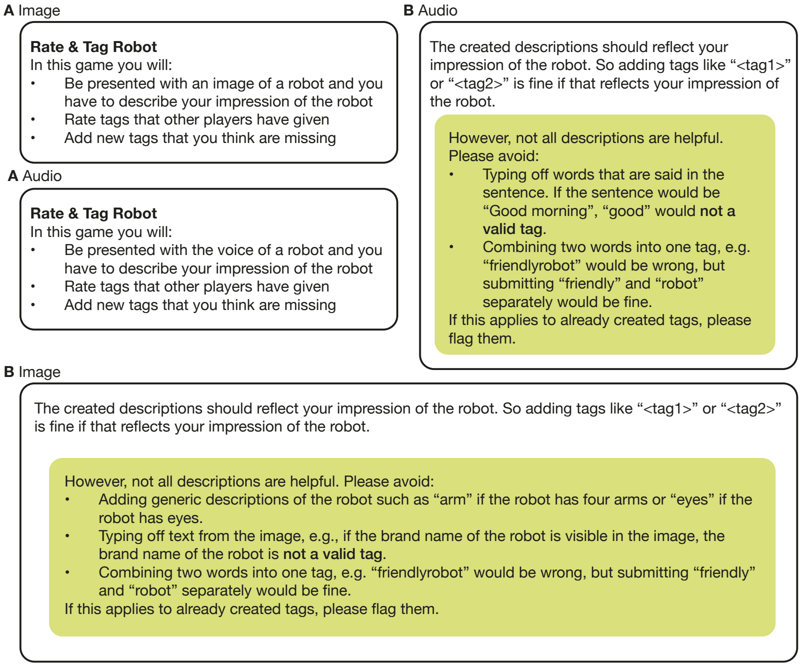
E.1 Instructions

The experiments proceeded as follows: upon completion of the consent form and the pre-screening tasks, participants received instructions regarding the main experiment (Figure S13–S14).

"<tag1>" or "<tag2>" are randomly selected from the following terms which were commonly used in a previous pilot:

- friendly,
- cute,
- functional,
- weird,
- humanlike,
- creepy,
- strange,
- odd,
- scary,
- unsettling,
- uncanny,
- powerful

A screenshot of the task is shown in Figure S15.



A Image

Rate & Tag Robot
In this game you will:
- Be presented with an image of a robot and you have to describe your impression of the robot
- Rate tags that other players have given
- Add new tags that you think are missing

A Audio

Rate & Tag Robot
In this game you will:
- Be presented with the voice of a robot and you have to describe your impression of the robot
- Rate tags that other players have given
- Add new tags that you think are missing

B Audio

The created descriptions should reflect your impression of the robot. So adding tags like "<tag1>" or "<tag2>" is fine if that reflects your impression of the robot.

However, not all descriptions are helpful.
Please avoid:
- Typing off words that are said in the sentence. If the sentence would be "Good morning", "good" would **not a valid tag.**
- Combining two words into one tag, e.g. "friendlyrobot" would be wrong, but submitting "friendly" and "robot" separately would be fine.

If this applies to already created tags, please flag them.

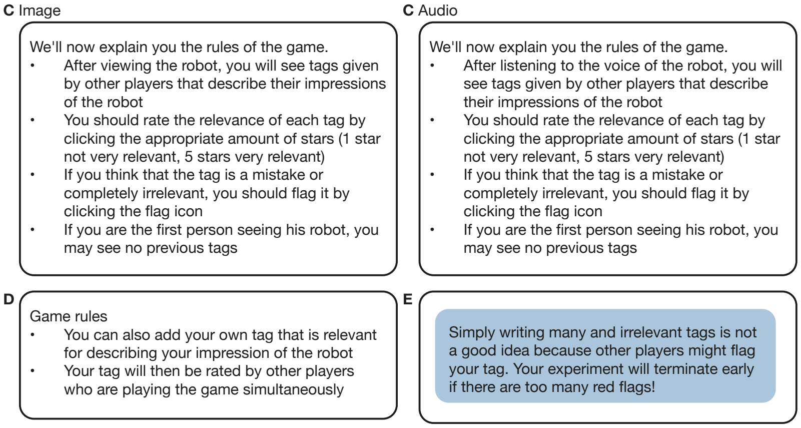
B Image

The created descriptions should reflect your impression of the robot. So adding tags like "<tag1>" or "<tag2>" is fine if that reflects your impression of the robot.

However, not all descriptions are helpful. Please avoid:
- Adding generic descriptions of the robot such as "arm" if the robot has four arms or "eyes" if the robot has eyes.
- Typing off text from the image, e.g., if the brand name of the robot is visible in the image, the brand name of the robot is **not a valid tag**.
- Combining two words into one tag, e.g. "friendlyrobot" would be wrong, but submitting "friendly" and "robot" separately would be fine.

If this applies to already created tags, please flag them.

Figure S13: Instructions for STEP experiment part 1/2.



C Image

We'll now explain you the rules of the game.
- After viewing the robot, you will see tags given by other players that describe their impressions of the robot
- You should rate the relevance of each tag by clicking the appropriate amount of stars (1 star not very relevant, 5 stars very relevant)
- If you think that the tag is a mistake or completely irrelevant, you should flag it by clicking the flag icon
- If you are the first person seeing his robot, you may see no previous tags

C Audio

We'll now explain you the rules of the game.
- After listening to the voice of the robot, you will see tags given by other players that describe their impressions of the robot
- You should rate the relevance of each tag by clicking the appropriate amount of stars (1 star not very relevant, 5 stars very relevant)
- If you think that the tag is a mistake or completely irrelevant, you should flag it by clicking the flag icon
- If you are the first person seeing his robot, you may see no previous tags

D

Game rules
- You can also add your own tag that is relevant for describing your impression of the robot
- Your tag will then be rated by other players who are playing the game simultaneously

E

Simply writing many and irrelevant tags is not a good idea because other players might flag your tag. Your experiment will terminate early if there are too many red flags!

Figure S14: Instructions for STEP experiment part 2/2.

E.2 STEP with implicit bias training

There is considerable overlap between the tags obtained in both STEP-Tag experiments. As shown in Figure S16, there is a strong correlation ($r = .78$) between the frequency of the tags. Indicating that the same tags were used to describe the robots. If the data of the STEP experiment after the awareness training had been used to compile the list of 40 terms, it would have led to the replacement of only a single term. The term "functional" would have been replaced by "modern". The results show that the awareness training had very little effect on the collected data.

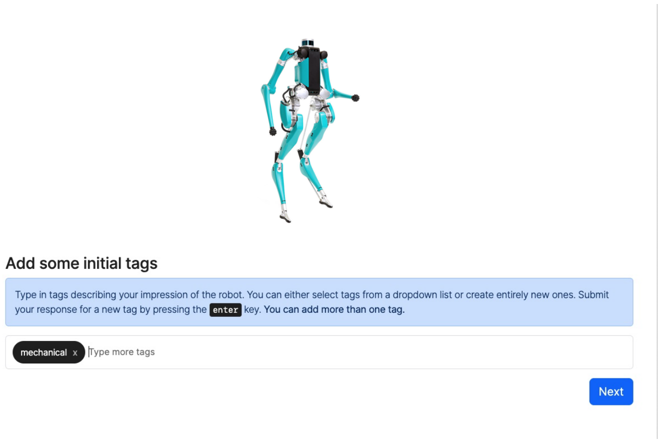
F RATE ROBOTS

F.1 Labels

The 40 used dimensions and their sources are listed in Supplementary Table S8.

F.2 Instructions

The instructions are depicted in Figure S17, a screenshot of the task is shown in Figure S18.



Add some initial tags

Type in tags describing your impression of the robot. You can either select tags from a dropdown list or create entirely new ones. Submit your response for a new tag by pressing the `enter` key. **You can add more than one tag.**

Figure S15: Example trial in the image STEP experiment.

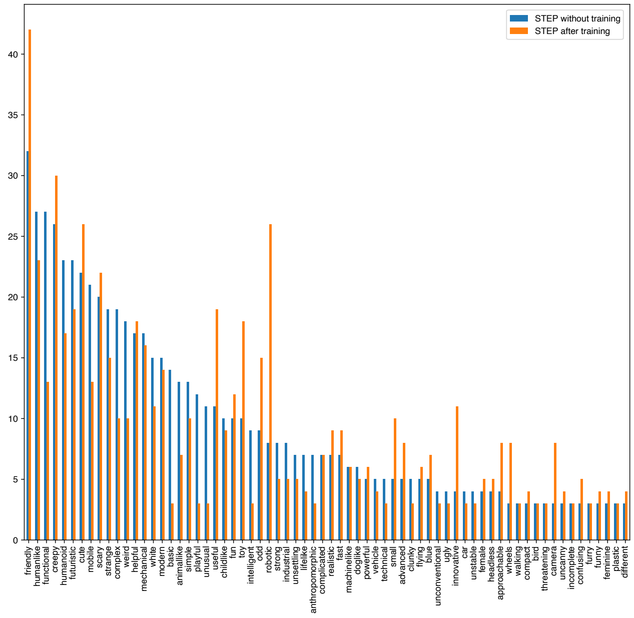


Figure S16: Frequency of terms with or without implicit bias training.

F.3 Consistency between STEP-Tag and dense rating

We computed the correlation between the STEP-Tag ratings and the dense ratings. For most dimensions, the value on the main diagonal is relatively large compared with other values, as shown in Figure S19. This suggests that dimensions were rated similarly across the two experiments. Furthermore, we found block-like structure for words with a similar meaning (e.g., female and feminine) or words with opposite meanings (e.g., male and female), indicating semantic clusters of terms. Some of these clusters also extend beyond the exact meaning or antonym, , for example "cute" (STEP-Tag) is not only highly correlated with "cute" (Dense), but also with "friendly" and "playful". Interestingly, the strength of the diagonal differs across the two modalities. For example, "scary" has a rather strong correlation in the visual modality, but is weaker in the voice modality. Generally, the correlation seems strongest for dimensions that are most salient in that modality: For example, the biological sex of a speaker and the clarity of their voice are salient, and features such as "humanlike", "animallike", or "friendly" have clear visual cues.

F.4 Consistency across modalities

Figure S20 shows the same data as in Figure 6D but now the dimensions are sorted by the strength of the diagonal.

F.5 Generalizability of the findings

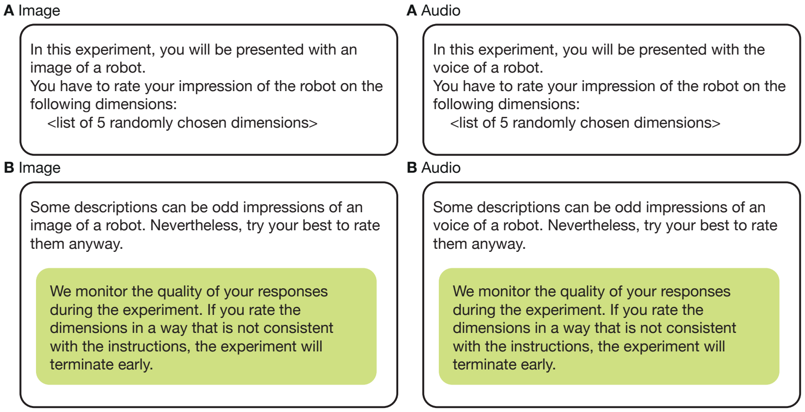
To assess the generalizability of the findings, we run the dense rating on 175 new robot images (see Supplementary Materials C.2) and on the initial random voices (i.e., iteration 0). We show in Figure S21, that the obtained correlation matrices strongly correlate with the initial correlation matrices: $r = 0.85$ for the image and $r = 0.91$ for the voice modality. These findings indicate that the obtained correlations across the terms are robust across databases.

F.6 Acoustic correlates

Figure S22 shows the correlations between the voice features and the perceptual dimensions. The first and third voice dimensions are correlated with an older male voice. The other latent voice dimensions do not correlate strongly with the 40 dimensions. Speaking speed correlates with "fast", "unclear", "playful", and "intelligent". All acoustic effects strongly correlate with "artificial", "robotic", "strange", and "unnatural".

F.7 Dense rating with implicit bias training

As shown in Figure S23, the implicit bias awareness training barely changed the correlations across terms as indicated by the high correlation across the upper triangles without diagonals ($r = .91$).



A Image

In this experiment, you will be presented with an image of a robot.
You have to rate your impression of the robot on the following dimensions:
<list of 5 randomly chosen dimensions>

A Audio

In this experiment, you will be presented with the voice of a robot.
You have to rate your impression of the robot on the following dimensions:
<list of 5 randomly chosen dimensions>

B Image

Some descriptions can be odd impressions of an image of a robot. Nevertheless, try your best to rate them anyway.

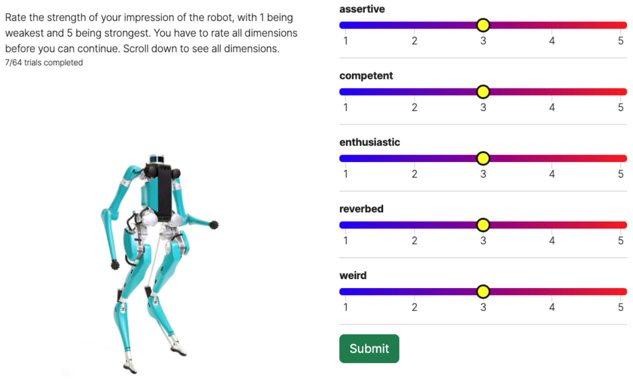
We monitor the quality of your responses during the experiment. If you rate the dimensions in a way that is not consistent with the instructions, the experiment will terminate early.

B Audio

Some descriptions can be odd impressions of an voice of a robot. Nevertheless, try your best to rate them anyway.

We monitor the quality of your responses during the experiment. If you rate the dimensions in a way that is not consistent with the instructions, the experiment will terminate early.

Figure S17: Instructions for dense rating experiment.



Rate the strength of your impression of the robot, with 1 being weakest and 5 being strongest. You have to rate all dimensions before you can continue. Scroll down to see all dimensions.
7/64 trials completed

assertive: 3
competent: 3
enthusiastic: 3
reverbed: 3
weird: 3

Figure S18: Example trial in the dense rating experiment.

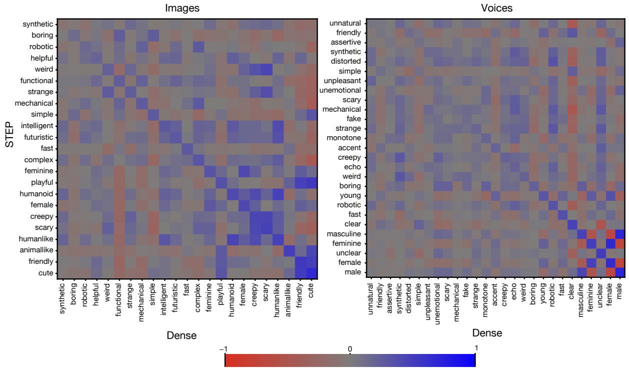


Figure S19: Consistency between STEP-Tag and dense rating. Correlation matrices are sorted by the strength of the diagonal.

While participants closely followed the implicit bias training (correct comprehension questions), it barely influenced the correlations across the terms.

F.8 Replace dense rating by deep learning model

To investigate if we can replace the dense rating procedure by a deep learning model, we provide two analyses on the pretrained CLIP model.[20]

In the first analysis, we compute the correlation between the cosine similarity computed on the dense image rating (which was used to predict a voice) and the cosine similarity of the image embedding. For both the old (r = .58) and the new set of 175 images (r = .51) we found a moderate correlation between the upper triangles of both cosine similarity matrices. This indicates that CLIP provides a fair proxy for the perceived similarity of robots.

In the second analysis, we use CLIP to do the dense rating. For each image, we obtain a logit value for each of the 40 dimensions. In Figure S24 we show the correlations across the 40 dimensions. Generally, the correlations across the dimensions are high in CLIP. There is a similarly strong correlation across the CLIP results across datasets (r = .85) compared to the dense rating results (r = .87). The results show that the correlational structure across the terms is

[20] https://github.com/openai/CLIP

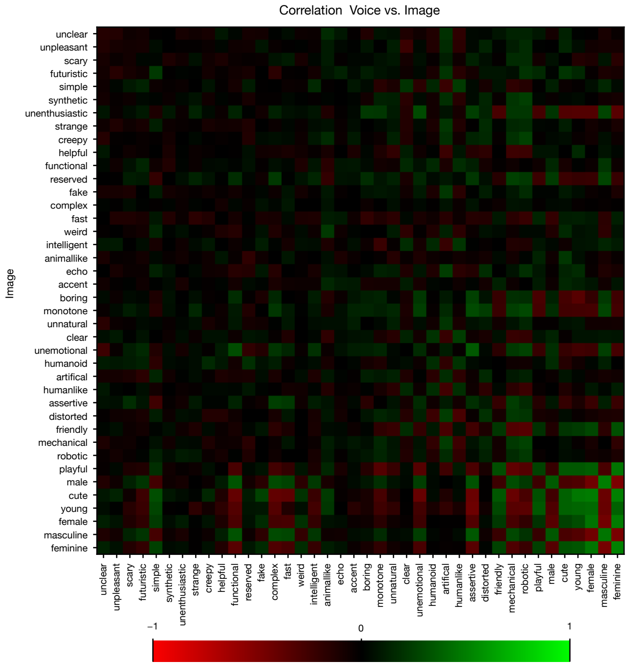


Figure S20: Correlation matrix between the ratings in the image and voice modality. An interactive version of this plot is available at: https://robotvoice.s3.amazonaws.com/compare.html.

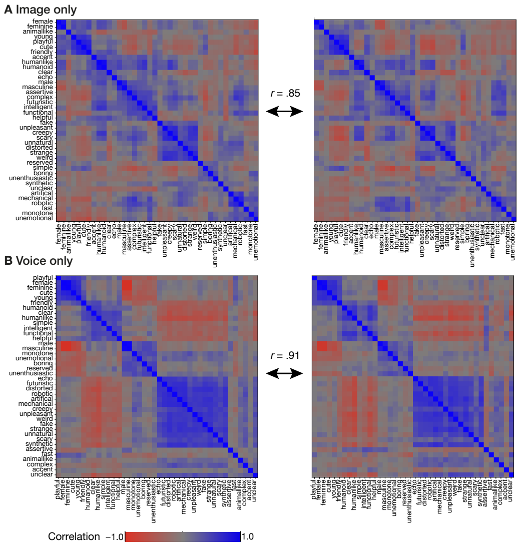


Figure S21: Replication of dense rating experiments on 175 new robot images (A) and on the initial random voices (B).

consistent across the datasets, but varies greatly between CLIP and the human dense rating.

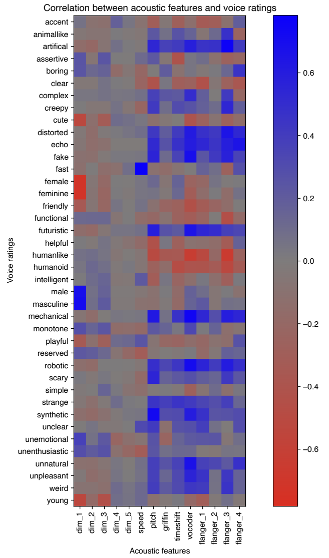


Figure S22: Correlations between voice features and perceptual dimensions.

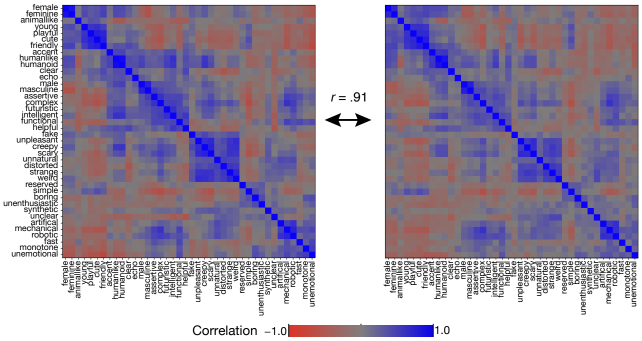


Figure S23: Correlation across correlation matrices in dense image rating experiment with or without implicit bias training.

G PREDICTION

G.1 Instructions

The instructions are identical to the GSP validation experiment (Section D.3).

G.2 Prediction per participant

To assess if the prediction result can also be found in single participants, we z-scored all ratings by each participant. We then computed the mean per participant and condition which are depicted as thin lines in Figure S25. As shown in Figure S25, most participants show the trend consistent with overall mean (thick line).

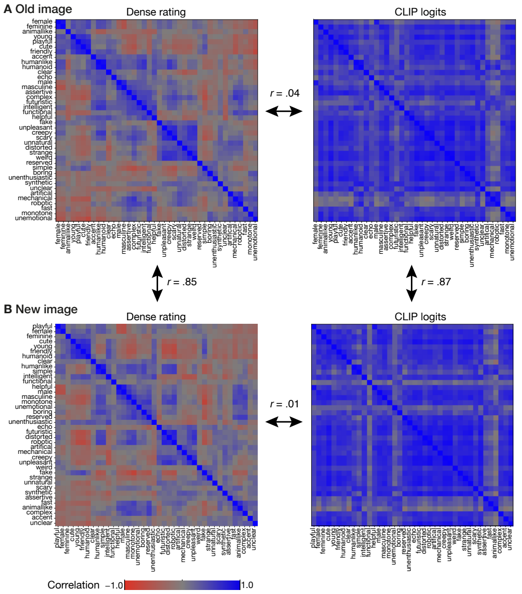


Figure S24: Correlation across 40 dimensions for human dense labeling and CLIP.

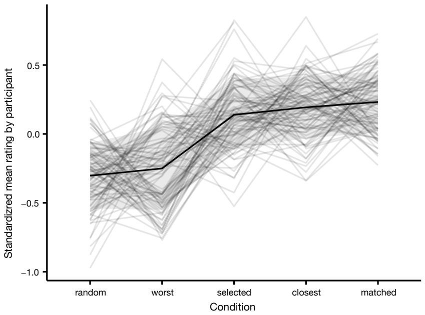


Figure S25: Prediction per participant. Mean standardized rating by participants and condition. Single lines depict single participants. The dark black line is the average across participants. The shaded area is the standard deviation across participants.

G.3 Factor analysis

To predict the closest robot based on perceptual dimensions, we performed a factor analysis on the 40-dimensional image ratings. Of the 40 dimensions, 39 are correlated at least 0.3 with at least one other feature, suggesting reasonable factorability. The Kaiser–Meyer–Olkin measure of sampling adequacy is 0.85, and Bartlett's test of sphericity is significant (5013.3, $p < 0.001$). Therefore, we applied factor analysis with Varimax (orthogonal) rotation.

We selected a seven-factor solution because the first seven eigenvalues are > 1 (Figure S26A). The factors explain 21%, 21%, 20%, 10%, 10%, 7%, and 2% of the variance (91% in total). Factor 1, "humanlike", mainly loads on humanlike and humanoid (see Figure S26B for the loading plot, the factor name is given by the dimension with the strongest loading). Factor 2, "cute", loads mainly on cute, friendly, playful, and young. Factor 3, "creepy", loads on creepy, distorted, scary, strange, unpleasant, and weird. Factor 4, "gender", positively loads on male and negatively on female. Factor 5, "natural" negatively loads on artificial, mechanical, and robotic. Factor 6, "fast", mildly loads on animallike, assertive, and fast. Factor 7, "functional", mildly loads on functional, helpful, and intelligent.

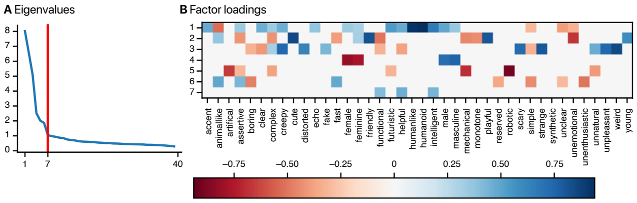


Figure S26: Factor analysis. A Eigenvalues plot. B Factor loading plot. Weak loadings (< 0.3) are omitted for the readability of the figure.

source(s)	adjective	in STEP	
GS, VA	friendly	□	▽
GS, VA'	mechanical	□	▽
AD, TIPI, VA'	simple	□	▽
TIPI, BFI-10	reserved		▽
IASR-B5, FFM, VA	assertive		▽
GS, VA	fake		▽
GS	unpleasant		▽
GS	artifical		▽
VA	synthetic		▽
PCPS*, HRG	unemotional		▽
AD	clear		▽
TIPI*	unenthusiastic		▽
AD'	boring		▽
GS*	unnatural		▽
AD*	unclear		▽
○	masculine		▽
○	male		▽
○	young		▽
○	feminine		▽
○	female		▽
VA	playful	□	
GS, VA	intelligent	□	
VA	helpful	□	
GS	humanlike	□	
TIPI	complex	□	
○	animallike	□	

markers:

□	in STEP-images
▽	in STEP-voices
*	antonym in source
'	synonym in source
○	added after pilot study

sources:

IASR-B5 [S81]	HRG [S28]	AD [S26]
FFM [S48]	PCPS [S15]	AD-BG [S25]
TIPI [S21]	GS [S8]	
BFI-10 [S66]	VA [S88]	

Table S4: List of attributes compiled from literature and pilot study. Part 1/4, showing those of the original 260 attributes that were also used by participants in the labeling task.

source(s)	adjective
TIPI	uncreative
AD	conservative
TIPI	open to experience
IASR-B5	unphilosophical
IASR-B5	unreflective
AD, TIPI	conventional
FFM*, AD, BFI-10*	unimaginative
AD, TIPI*, VA	creative
IASR-B5*, BFI-10*, VA	artistic
IASR-B5, FFM, BFI-10’	imaginative
IASR-B5, TIPI*	unconventional
IASR-B5, FFM*	unreliable
AD	innovative
IASR-B5	questioning
IASR-B5	philosophical
IASR-B5	reflective
FFM	curious
AD	original
VA	joyful
IASR-B5	broad-minded
BFI-10, VA	lazy
VA	principled
VA	reckless
AD	predictable
GS	irresponsible
TIPI	careless
IASR-B5, FFM*, TIPI, VA*	disorganized
IASR-B5, FFM*	inefficient
IASR-B5	unsystematic
IASR-B5*, FFM*, VA	superficial
IASR-B5	undisciplined
VA	messy
BFI-10*	diligent
IASR-B5, FFM	reliable
AD	unpredictable
GS	responsible
IASR-B5, FFM, TIPI*, VA	organized
IASR-B5	orderly
IASR-B5, FFM	efficient
IASR-B5	systematic
IASR-B5, FFM, BFI-10’, VA	thorough
IASR-B5, VA*	tidy
TIPI	extroverted
TIPI	quiet
IASR-B5	timid
IASR-B5, VA*	forceless
IASR-B5	meek
FFM	talkative
FFM’, VA	expressive
FFM	active
IASR-B5	dominant
VA	powerful
FFM, BFI-10, TIPI*	outgoing
IASR-B5, VA	forceful
IASR-B5	firm
FFM	energetic
FFM, VA	enthusiastic
VA	agreeable
FFM*, BFI-10*, VA	distrustful
VA	detached
GS, VA*	unfriendly
IASR-B5	uncharitable
IASR-B5	soft-hearted
IASR-B5*, FFM*, GS, VA*	unkind
IASR-B5, VA	cruel
FFM*, VA	stingy
IASR-B5	ruthless
GS	pleasant
TIPI*, VA	peaceful
FFM, BFI-10, VA’	trusting
VA	benevolent
VA	affectionate
VA	respectful
IASR-B5, FFM, TIPI	sympathetic
IASR-B5	charitable
IASR-B5	iron-hearted
IASR-B5*, FFM, TIPI, VA	warm
IASR-B5, FFM, GS, VA	kind

markers:

*	antonym in source
’	synonym in source
○	added after pilot study

sources:

IASR-B5 [S81]	AD [S26]
FFM [S48]	GS [S8]
TIPI [S21]	VA [S88]
BFI-10 [S66]	

Table S5: List of attributes compiled from literature and pilot study. Part 2/4, showing a subset of the original 260 attributes that was not used by participants in the labeling task.

source(s)	adjective
IASR-B5	tender
FFM	appreciative
FFM	forgiving
FFM	generous
BFI-10	sociable
IASR-B5*, FFM, TIPI*	unstable
IASR-B5, FFM*, TIPI	stable
IASR-B5, BFI-10, VA	nervous
VA	temperamental
FFM	impulsive
IASR-B5, FFM	worrying
IASR-B5, FFM	tense
IASR-B5, FFM*, TIPI*, VA*	unanxious
IASR-B5’, VA	excitable
FFM	thin-skinned
IASR-B5*, VA	moody
FFM	touchy
IASR-B5, TIPI, VA	calm
VA	stoic
FFM*, VA	deliberate
IASR-B5, FFM*	unworrying
IASR-B5, BFI-10, VA	relaxed
IASR-B5, FFM, TIPI, VA	anxious
GS	incompetent
GS	competent
GS, VA	ignorant
VA	dumb
GS	knowledgeable
VA	useful
GS	natural
GS	machinelike
GS	unconscious
GS, VA	dead
GS	stagnant
GS	inert
GS	conscious
GS, VA*	alive
GS	lively
GS	organic
GS, VA	interactive
GS, VA	responsive
AD	not presentable
AD	unstylish
AD	confusing
AD	cumbersome
AD	complicated
VA	soothing
AD	presentable
AD	valuable
AD	stylish
AD	direct
AD	engaging
GS	moving rigidly
GS	lifelike
GS	moving elegantly
VA	flexible
GS	apathetic
GS, VA*	unintelligent
GS	foolish
GS	sensible
GS	dislike
GS	awful
GS	like
GS	nice
BFI-10’, VA	fault-finding
FFM*, TIPI	critical
TIPI	quarrelsome
FFM, TIPI, VA	dependable
IASR-B5, FFM, TIPI, VA	self-disciplined
GS	agitated
GS	surprised
GS	quiescent
PCPS’, HRG	thoughtful
PCPS	inattentive
HRG	cautious
HRG	reasonable
PCPS	honest
HRG	weak
PCPS	arrogant
HRG	uncooperative

markers:

*	antonym in source
’	synonym in source
○	added after pilot study

sources:

IASR-B5 [S81]
FFM [S48]
TIPI [S21]
BFI-10 [S66]
AD [S26]
GS [S8]
VA [S88]

Table S6: List of attributes compiled from literature and pilot study. Part 3/4, showing a subset of the original 260 attributes that was not used by participants in the labeling task.

source(s)	adjective
HRG	impolite
HRG	cooperative
HRG	polite
PCPS	merciful
PCPS	emotional
AD-BG	ugly
AD-BG	beautiful
TIPI*, VA*	closed-minded
BFI-10’, VA*	unartistic
TIPI’	open-minded
VA*	shallow
IASR-B5*	unquestioning
IASR-B5*	narrow-minded
FFM*	uncurious
FFM*	unoriginal
VA*	serious
VA*	unprincipled
IASR-B5, AD	impractical
IASR-B5’	disorderly
TIPI*	careful
IASR-B5*, AD	practical
IASR-B5’, FFM’, VA’	disciplined
TIPI*	introverted
FFM*, VA*	expressionless
FFM*	passive
IASR-B5*, FFM*, VA*	non-assertive
IASR-B5*	submissive
VA*	powerless
FFM*	non-energetic
IASR-B5*	bold
VA*	disagreeable
VA*	belligerent
VA*	malevolent
VA*	disrespectful
IASR-B5*, FFM*, TIPI*	unsympathetic
IASR-B5’, TIPI*	cold
FFM*	unappreciative
FFM*	unforgiving
IASR-B5*, VA*	non-excitable
FFM*	thick-skinned
VA*	unhelpful
VA*	useless
GS’, VA*	unresponsive
GS’	agitating
AD’	worthless
GS’	calming
FFM*, BFI-10*	reclusive
BFI-10*	unsociable
FFM’	uncritical
FFM*, TIPI*, VA*	undependable
TIPI’	upset
TIPI’	loud
HRG*	unreasonable
PCPS*	dishonest
PCPS*	attentive
PCPS*	merciless
GS’	intimidating
GS’	upsetting
GS’	reassuring
○	small
○	tiny
○	old
○	big
○	tall
○	distant
○	involved
○	changing
○	constant
○	repulsive
○	unattractive
○	attractive
○	inelegant
○	uninteresting
○	elegant
○	interesting
○	uncomfortable

markers:

*	antonym in source
’	synonym in source
○	added after pilot study

sources:

IASR-B5 [S81]	AD [S26]
FFM [S48]	GS [S8]
TIPI [S21]	VA [S88]
BFI-10 [S66]	

Table S7: List of attributes compiled from literature and pilot study. Part 4/4, showing a subset of the original 260 attributes that was not used by participants in the labeling task.

source(s)	adjective	in STEP	
GS, VA	friendly	□	▽
GS, VA'	mechanical	□	▽
AD, TIPI, VA'	simple	□	▽
GS'	robotic	□	▽
	creepy	□	▽
	weird	□	▽
VA	playful	□	
GS, VA	intelligent	□	
VA	helpful	□	
GS	humanlike	□	
TIPI	complex	□	
GS'	humanoid	□	
GS'	scary	□	
○	animallike	□	
○	cute	□	
	strange	□	
	futuristic	□	
	functional	□	

source(s)	adjective	in STEP
TIPI, BFI-10	reserved	▽
IASR-B5, FFM, VA	assertive	▽
GS	unpleasant	▽
PCPS*, HRG	unemotional	▽
GS	artificial	▽
VA	synthetic	▽
AD	clear	▽
GS, VA	fake	▽
TIPI*	unenthusiastic	▽
GS*	unnatural	▽
AD*	unclear	▽
AD'	boring	▽
○	masculine	▽
○	male	▽
○	young	▽
○	feminine	▽
○	female	▽
	echo	▽
	accent	▽
	distorted	▽
	fast	▽
	monotone	▽

markers:

□	in STEP-images
▽	in STEP-voices
*	antonym in source
'	synonym in source
○	added after pilot study

sources:

IASR-B5 [S81]	HRG [S28]	AD [S26]
FFM [S48]	PCPS [S15]	AD-BG [S25]
TIPI [S21]	GS [S8]	
BFI-10 [S66]	VA [S88]	

Table S8: The 40 labels selected for rating the robots' images and the voices. Whenever possible, references that confirm them were added for those that were not in the original list of 260 attributes.

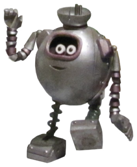
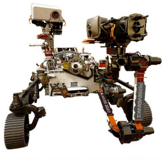
	Name: Bärbot **Creator:** University of Augsburg **Image Credits:** photo taken and edited by authors
	**Name:** Schlupp **Creator:** Augsburger Puppenkiste **Image Credits:** photo taken and edited by authors
	Name: RoboKind R50 Alice **Creator:** Hanson Robotics **Image Credits:** photo taken and edited by authors
	Name: Digit **Creator:** Agility Robotics **Image Credits:** Agility Robotics (written approval)
	Name: Daisy **Creator:** HEBI robotics **Image Credits:** HEBI robotics (written approval)
	Name: Valkyrie **Creator:** NASA **Image Credits:** NASA (not subject to copyright for non-commercial use)
	**Name:** Perseverance **Creator:** NASA **Image Credits:** NASA (not subject to copyright for non-commercial use)

Table S9: Copyright statement of robot images used in the figures of the manuscript (1/2).

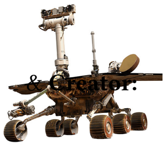
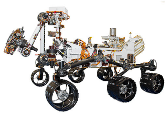
 & Creator:	**Name:** NASA **Image Credits:**	Spirit & Opportunity NASA (not subject to copyright for non-commercial use)
	**Name:** **Creator:** **Image Credits:**	Curiosity NASA NASA (not subject to copyright for non-commercial use)
	Name: **Creator:** **Image Credits:**	Robonaut 2 NASA NASA (not subject to copyright for non-commercial use)

Table S10: Copyright statement of robot images used in the figures of the manuscript (2/2).