# The Effect of Audio-Visual Smiles on Social Influence in a Cooperative Human–Agent Interaction Task

ILARIA TORRE, KTH Royal Institute of Technology and Trinity College Dublin
EMMA CARRIGAN, Trinity College Dublin
KATARINA DOMIJAN, Maynooth University
RACHEL MCDONNELL and NAOMI HARTE, Trinity College Dublin and ADAPT Research Centre

Emotional expressivity is essential for human interactions, informing both perception and decision-making. Here, we examine whether creating an audio-visual emotional channel mismatch influences decision-making in a cooperative task with a virtual character. We created a virtual character that was either congruent in its emotional expression (smiling in the face and voice) or incongruent (smiling in only one channel). People (N = 98) evaluated the character in terms of valence and arousal in an online study; then, visitors in a museum played the "lunar survival task" with the character over three experiments (N = 597, 78, 101, respectively). Exploratory results suggest that multi-modal expressions are perceived, and reacted upon, differently than unimodal expressions, supporting previous theories of audio-visual integration.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *Empirical studies in collaborative and social computing*; • **Applied computing** → *Psychology;*

Additional Key Words and Phrases: Multimodal emotional expression, artificial agent, social influence, smiling

**ACM Reference format:**
Ilaria Torre, Emma Carrigan, Katarina Domijan, Rachel McDonnell, and Naomi Harte. 2021. The Effect of Audio-Visual Smiles on Social Influence in a Cooperative Human–Agent Interaction Task. *ACM Trans. Comput.-Hum. Interact.* 28, 6, Article 44 (November 2021), 38 pages.
https://doi.org/10.1145/3469232

## 1 INTRODUCTION

Emotions are an essential aspect of social interaction, and as humans, we have evolved to accurately signal and perceive them in their different modalities, including facial cues [41], vocal cues [104], or bodily cues [61]. Emotional expressivity is also connected with positive personality traits,

The research was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 713567, and by the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2016) and is co-funded by the European Regional Development Fund. The second author received funding from the Science Foundation Ireland, Game Face (13/CDA/2135) project.
Authors' addresses: I. Torre, KTH Royal Institute of Technology, Stockholm 11428, Sweden, Trinity College Dublin, Dublin D 02 PN40, Ireland; email: ilariat@kth.se; E. Carrigan, Trinity College Dublin, Dublin D 02 PN40, Ireland; email: carrige@tcd.ie; K. Domijan, Maynooth University, Maynooth Co. Kildare, Ireland; email: katarina.domijan@mu.ie; R. McDonnell and N. Harte, Trinity College Dublin, ADAPT Research Centre, Dublin 2, Ireland; emails: {ramcdonn, nharte}@tcd.ie.

1073-0516/2021/11-ART44 $15.00
https://doi.org/10.1145/3469232

including cooperation and trust [12, 35, 72, 106], which are fundamental to a functioning society [7]. While emotional expressions have been studied extensively in the context of human perception and behaviour, the literature on multi-modal expression in machines is comparably scarce [31, 73]. However, as machines with "faces", "voices", and "bodies" increasingly populate our social space, studying the effect of machine characteristics—such as emotional expression—on human decision-making is paramount to ensure cooperation between humans and machines. For example, would people rather trust a happy or a neutral navigation system? And what would be the pre-dominant channel in the perception of the emotion expressed by an artificial agent?

Here, we present three experiments using the "lunar survival task" (Section 3.1), where participants cooperate with a virtual character. This task served as a proxy for measuring the willingness of people to accept the virtual character's recommendations with regards to which items would be most important for survival in a disaster scenario. We manipulated the vocal and facial emotional expressions of the character, so that it was either smiling or neutral, in all possible combinations. In this way, we wanted to examine the effect of different combinations of emotional expressivity on social decision-making. The experiments were conducted in a public setting in a museum adjacent to Trinity College Dublin, called the Science Gallery, where thousands of visitors interacted with our virtual agent. Preliminary analyses on a subset of participants were previously published in Torre et al. [114]; here, we report the complete analyses on the 835 participants who gave their consent to retain their data, plus an online study evaluating the used stimuli in terms of valence and arousal.

We first explore related work in the following section; then we detail the experimental setup in Section 3; in Sections 4, 5, 6, and 7 we present the results from the crowdsourced evaluation and the three museum experiments; in Section 8 we evaluate the results across the different experiments and highlight the relevance of our findings; and we conclude briefly in Section 9.

## 2 RELATED WORK

The function of emotions has been narrowed down to two main components: (1) to evaluate an external or internal event [103]. For example, Simon [108] defines emotions as responses to a sudden, intense stimulus from the environment that creates arousal in the autonomic nervous system. These responses are akin to an "interrupt system" which can set aside ongoing processes when real-time needs of high priority are encountered. (2) To communicate this evaluation to others, for example, in order to elicit a desired behaviour (e.g., an infant crying for attention) [69]. Conversely, emotional expressions can be used by on-lookers to draw situational information [56] and to inform decision-making [58]. This has been shown even for Human–Machine Interaction: people cooperate more with an artificial agent if this shows an emotional expression that signals its intention to be cooperative [28].

Here, it is worth making a distinction between emotions and emotional expressions: while the former indicate what someone is actually feeling, the latter indicate how someone is displaying an emotion, for example, by making certain facial expressions or speaking with affective prosody [22]. This implies that emotional expressions do not necessarily reflect a real emotion, but they can be deliberate, at least to a certain extent [39, 51].

### 2.1 Emotional Expressivity in Human–Machine Interaction

Emotions are of interest to the Human–Machine Interaction community: a machine that is able to accurately recognise how a user is feeling could react appropriately, thus making the interaction more natural and successful. Recently, socio-functional theories of emotion have also sparked research in another direction, namely studying how people respond to machines that display an emotional expression [4, 27, 28, 60, 115]. Some studies have found that an artificial agent displaying

an emotional expression is perceived more positively than a neutral one. For example, Elkins and Derrick [42] found that people rated smiling embodied conversational agents as more trustworthy than non-smiling ones; smiling also led to trusting avatars and robots more [62, 76]. Robots exhibiting an emotional expression were also "more fun to play with" in Leite et al. [66]. This might be because displaying a positive emotion generally leads to attributions of other positive traits, in a typical "halo effect" [65, 89, 106]. Additionally, since it has been pointed out that individuals who are particularly expressive might be less adept at disguising their emotions [30], it has been suggested that emotional expressions might be a sign of commitment to cooperative behaviour [47, 106]. As a consequence, individuals who display emotions, even negative ones, might be seen as less deceptive and more trustworthy [12, 106].

Environmental factors are presumed to be one of the additional sources that influence people's first impressions of an individual—or an artificial agent [cf. 53]. Specifically, an emotional expression is interpreted in context, and a mismatch between expression and behaviour can lead to negative reactions. For example, Antos et al. [4] found that, in a negotiation game, participants tended to select as partners computer agents that displayed emotions congruent with their actions. These agents were also perceived as more trustworthy than agents whose emotional expression and action strategy did not match, even though the strategy itself was the same. In Khooshabeh et al. [59], people played a negotiation game where they had to sell phone plans to an avatar, which had a mismatched facial expression (happy/angry) and negotiation strategy (tough/soft). They found that people felt more threat—as indexed by lower cardiac output—when the avatar had an angry expression but soft strategy, and similarly people looked more at the avatar's face in this condition. It might be that this combination sparks a "reaction to the unknown", so people try to look at the primary source of emotional expression (in this case, the face) to make sense of what is happening and obtain more information [cf. 117, 118]. These findings provide evidence for the "**Emotions As Social Information**" (**EASI**) model [118], which posits that emotions are used to make sense of ambiguous situations, and that their effect depends on the context in which the interaction takes place, being specifically mediated by its cooperative or competitive nature. Thus, displaying a positive emotion, such as happiness, in a cooperative context will reinforce the parties' belief that everyone is gaining, and will elicit more cooperative behaviours. By contrast, displaying a negative emotion, such as anger, in a cooperative context will hinder future cooperative behaviours.

However, the EASI model does not account for other evidence suggesting that positive emotional expressions have a mitigating effect on negative behaviour. For example, Mieth et al. [82] found that smiling counterparts were punished less harshly than neutral ones when they defected in a trust game with an option for monetary punishment. Similarly, participants in a simulated trial scenario gave less harsh sentences to people who appeared to be smiling in photographs, while perceiving the smiling transgressors as equally guilty as the non-smiling ones [63]. Recently, artificial agents that spoke with a happy-sounding voice were trusted more than neutral ones in a trust game, even when they were behaving equally untrustworthily [115]. This difference might be explained by the fact that the experiments that found a mitigating smile effect showed participants the emotional expressions *before* they were required to provide a judgement or an action, while the experiments that support the EASI model showed the emotional expressions *after* an agent performed a certain action. Thus, smiling after exploiting a partner in a trust game could make the smiler be perceived more negatively, while smiling before exploiting a partner might still transmit hope that the next action will be cooperative [101].

If the perception of emotional expression depends on the context of the interaction, as these studies suggest, it would be interesting to see how conflicting information regarding the emotional expression also affects perception: if the audio and visual channels are communicating a different emotion, would people act upon the emotion they perceive from one of the two channels? Or

would they act based on the perception of a combined emotion, in a sort of "emotional McGurk effect" [43]?

### 2.2 Perception of Incongruent Audio-Visual Emotional Stimuli

Most research featuring expressive artificial agents featured a "congruent" emotion, that was either displayed in the voice and face, or only in one of the two channels. For example, the agents in Mathur and Reichling [76] and Leite et al. [66] were smiling only in the face, while in Read and Belpaeme [93], Scheutz et al. [105], Torre et al. [115] they were expressing an emotion only in the voice. This makes sense when the study focuses on people's reactions to an agent displaying a certain emotion, since people might find a channel mismatch confusing [23, 71]. Indeed, congruent audio and video information can facilitate perception and understanding, as in the case of lip-reading, while incongruent information can interfere with it, as in the case of the McGurk effect [17, 80].

However, mismatching the emotion expressed in the face and voice is a technique that can be used to learn more about channel pre-dominance in emotion perception: do we understand that someone is happy from their face—e.g., their smile—or their voice—e.g., their affective prosody? Studies using human sources, such as photographs of a happy person paired with audio of a sad person, have shown that people are more accurate at recognising an emotion when face and voice match, and, when they do not match, they extract more accurate information from the visual channel [25, 43, 111]. Additionally, Creed and Beale [23] found that emotions were perceived more strongly when expressed in one channel, despite other channels presenting a contrasting expression. Based on these results, it has been suggested that the audio and video channels express different components of emotions; specifically, the video channel mostly expresses the emotional valence (positive–negative) and the audio channel the emotional activation (high arousal–low arousal) [43, 54, 86].

Mower Provost and colleagues extended these results to expressive artificial agents: they mixed and matched an animated character's face expressing an emotion (angry, sad, neutral, and happy) and an actress's voice expressing the same emotions over semantically neutral utterances [85, 86]. For each participant, 50% of the stimuli were audio-visual, 25% video only and 25% audio only; participants rated them in terms of valence, activation and dominance using the **self-assessment manikins** (**SAM**) method (originally devised by Bradley and Lang [13]). Similarly to the studies using human sources, they found that the classification rate was highest for matching audio-visual stimuli. However, they also found that classification rate was also higher when there was audio only than video only; and, when faced with a conflicting emotional presentation, participants pre-dominantly attended to the vocal channel. This surprising result was explained in terms of naturalness mismatch: their animated character face was limited in terms of expressivity, while the voice came from a human actress, so people might have focused on the most reliable source of information in these experiments. Thus, it remains to be seen how an "emotional McGurk effect" influences emotional processing in artificial agents that are equally expressive in the face and voice.

### 2.3 Our Approach

These studies suggest that, when presented with conflicting information, people might tune to the pre-dominant channel to perform appraisal of the presented emotion. However, the question of what actions would be elicited by this conflicting information remains unanswered, as previous studies mostly focused on emotion *recognition*—whether using categorical emotion labels [25], or multi-dimensional continua [86]. However, asking participants to recognise an emotion is problematic, since providing emotion labels to choose from could bias them to choose a label that does

not fully represent their perception [9, 20, 45]. Furthermore, this approach risks missing out on subtle variabilities in the emotions continuum that a mismatched expression likely causes [cf. the McGurk effect 80], since people are typically "forced" to choose out of a set of pre-defined emotion labels (but see Mower et al. [85, 86] for a slightly different approach using the SAM method). Also, previous studies have mostly focused on emotion recognition from a human source, e.g., a human face and a human voice [25, 90], and it remains to be seen whether these results will extend to artificial sources. Humans cannot produce a certain emotion in the face and another in the voice, so this approach raises issues around the intrinsic validity of the results. On the other hand, communication channels can be operated independently in an artificial agent [e.g., 86, 95]. This would not only help study the often overlooked "big picture" of multi-modal emotional expressivity, but it would also have implications for artificial agent design, even if this entails a humanly impossible channel incongruency. As one of our goals as Human–Machine Interaction researchers is making interactions with machines work, nothing prevents us from designing a machine that expresses a mismatched emotion, if there is reason to do so. Does the emotional expression need to be present in both channels to increase the naturalness and success of the interaction? And if not, is this an effect of the combined efficacy of a mismatched emotional expression, or is it because one channel is pre-dominant over the other?

In the present work, we borrow methodologies from emotion integration theories [24, 25, 90] and apply them to study decision-making in Human–Agent Interaction [4, 27, 28, 59]. Specifically, we examine the social influence of a virtual character that combines vocal and facial emotional expressions in a matched and mismatched fashion. Our virtual character is created using state-of-the-art motion capture techniques, making its emotional displays realistic, thus advancing the knowledge from previous studies [85, 86].

### 2.4 Smiles

In our study, we focus on smiles. Smiling is a universal [81], multi-modal [38, 44, 113] emotional expression, which is visible in the face and audible from the voice [6, 40]. Smiling does not always express the same emotion: there are sad smiles, convenience smiles, Duchenne smiles, mocking smiles, and many others [e.g., 34, 39, 100]. However, different types of smiles have been shown to increase cooperation and trust [42, 94, 115]. After finding that smiling faces elicited positive trait attributions, Lau [65] argued that this could be due to a "halo effect": smiling is good, so people who smile must possess other good traits, too [101]. Similarly, Penton-Voak et al. [89] concluded that people with facial features that elicit attributions of agreeableness—one of the "Big Five" personality traits, related to interpersonal relations [79]—may be treated as more trustworthy and may consequently develop more agreeable personality characteristics. This has been referred to as the "what is beautiful is good" stereotype [33], and even "what is smiling is beautiful and good" [96]. By manipulating an artificial agent's smile, we want to shed light on the effect of this emotional expression on human behaviour.

### 2.5 The Survival Task as a Measure of Social Influence and Cooperative Behaviour

We used the "lunar survival task" to measure the avatar's social influence and participants' cooperative behaviour. In this paradigm, the interactants have to come to an agreement on the order of importance of a set of items in a simulated dangerous scenario, such as a moon crash. The survival task is more representative of choices we might need to make in our everyday lives (e.g., accepting someone's counsel) rather than abstract actions such as those in a trust game [10].

The survival task has different variants, desert and moon being the most widely used. It was originally devised to measure leadership and dominance in group settings (e.g., a group of employees), for example, by looking at who in the group tries to force his/her positions on the others

[52]. Since then, it has been used in a number of Human–Machine Interaction studies to measure persuasiveness in robots [1, 18, 112] and virtual characters [60, 87, 119]. People seem to be willing to accept recommendations from artificial agents in this task, either when the agent is a robot, a virtual character, or a computer [5, 87].

Using an artificial agent to disagree on the ranking order allows to measure how much people are willing to change their own positions and accept the agent's. This willingness to accept a different ranking highlights that several interpersonal processes are at work. First of all, participants' initial ranking in the task is entirely dictated by their knowledge on the situation at hand (e.g., space physics), and, in the absence of this, by common sense. Second, when the agent appears to the participants and provides its ranking, participants will form a first impression of the agent based on many different factors: agent-related factors (human-likeness, attractiveness, voice quality, rendering style, affective state, non-verbal behaviour, etc. [e.g., 16, 32, 122]), human-related factors (experience with artificial agents, attitudes towards artificial agents, perception of similarity, own affective state, etc. [e.g., 107]), and environment-related factors (e.g., "does the agent's presence in the moon crash scenario make the situation more positive or more negative?" "What is the nature of the relationship between me and the agent?" [e.g., 53]). These first impressions are basically immediate [8, 78], and will thus form before the agent provides the reasoning behind its ranking. Third, the plausibility of the agent's justification for its recommendations will be evaluated, and this evaluation is likely to be mediated by the previously formed first impressions [cf. 68, 116]. When participants make their second and final ranking, all these factors will be used to make a decision.

The "lunar survival task" is framed as a cooperative game, so participants are primed to believe that the agent might not want to intentionally provide wrong information. However, participants might deem to have more knowledge than the avatar, perhaps because they consider it a subordinate, rather than a peer [70]. It might also be that they have a strong opinion about the objects to rank—we tried to limit this possibility by asking players to rank objects that all had pros and cons. In any case, if participants end up accepting the agent's recommendation, it is likely due to the fact that they assume the agent to be knowledgeable, and they deem its message persuasive. As all objects have inherent positive and negative aspects, accepting the agent's suggestion shows that the agent's knowledge on the topic is deemed bigger than one's own, and that people have confidence in it. This confidence is an integrative part of trust and persuasion theories [11, 21]. Also, albeit simulated, the survival task is a risk scenario: participants risk loosing their (virtual) life if they pick the wrong items. In this sense, accepting the agents' recommendations also represents a willingness to take a risk, in case they turn out to be wrong. This willingness is also a component of trust [77].

Thus, the survival task encompasses multiple interpersonal phenomena—persuasion, trust, knowledgeableness, social status, and more. Here we choose to use the broad term "social influence", previously adopted by Artstein et al. [5], to describe the agent feature driving participants' reactions.

## 3 METHOD

We first ran an online study where we examined how people perceived our matched and mismatched audio-visual smiles in terms of valence and arousal. We then ran our three main experiments in a museum, where visitors could play the "Lunar survival task" with an artificial agent displaying these different types of smiles. Below, we first describe our implementation of the survival task and how the artificial agent's stimuli were created (Sections 3.1, 3.2, 3.3, and 3.4). The procedure and results of the four total experiments are then presented in Sections 4, 5, 6, and 7.

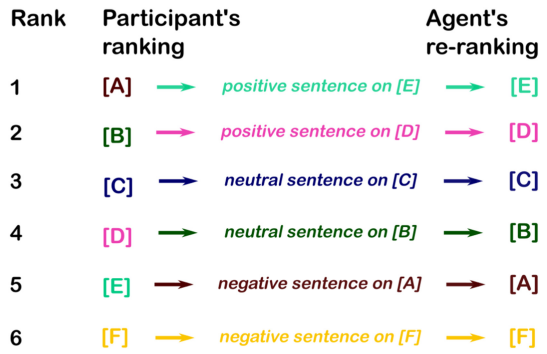


Fig. 1. Diagram showing the changes made by the virtual character to the participants' rankings of the objects (A–E in the diagram).

### 3.1 "Lunar Survival Task" Design

At the beginning of the game, participants are told that they have crash-landed on the surface of the moon with only 6 intact objects, and that their mother ship is located 200 miles away. They have to rank the 6 intact objects in order of importance for survival, to make a decision on which objects to carry with them during the journey. After they make the initial ranking, an avatar, which was originally meant to be their navigation assistant, appears and suggests a different ranking order (Figure 1). For example, the avatar places whichever object participants had placed at position 1 (most important), to position 5 (second least important); and so on. Thus, for all participants, the difference between their initial ranking and the artificial agent's ranking is constant (Kendal-Tau distance = 10). This is the same manipulation used in Moon [84] and Nass et al. [87], the first studies to employ the survival task in Human–Machine Interaction. However, other studies used an agent's ranking that was pre-determined independently of the participants' [e.g., 5, 74]; this was used to calculate a divergence score from the artificial agent before and after the agent's ranking. We chose to have full control over how much the agent diverged from the participants, so that this manipulation would be the same for all participants. After the avatar made these suggestions, participants could make their final decision, by providing a final ranking order. The comparison between the avatar's ranking and the participants' final ranking was the basis for our analysis of social influence. The main experimental procedure is shown in Figure 2.

The 6 objects were chosen from a longer list published by NASA [52]. Here, we differ from the design used in previous experiments [84, 87], where participants had to rank the full list of 15 objects. We wanted participants to be less likely to have strong opinions about objects that would obviously be very important on the moon (such as oxygen tanks) or very useless (such as a box of matches), so that they might be equally likely to demote or promote any item. So we eliminated the NASA top 2 and bottom 2 objects, and chose 6 objects between the remaining 11. These objects were: nylon rope, parachute silk, portable heating unit, milk tank, life raft, and receiver transmitter. As it was expected that participants would not be experts in space exploration, and as the original ranking of the objects was neither too high nor too low in the NASA list, we assumed that each object would start with the same probability of being chosen as most important [cf. 2, 3].

### 3.2 Utterance Preparation

We prepared a script encompassing all the utterances that the avatar could speak during the "lunar survival task". As the avatar's ranking depends on the participant's initial ranking (see Figure 1),


Fig. 2. Diagram showing the main experimental procedure. Participants make their initial ranking in (1); the virtual character makes its ranking in (2); participants make their final ranking in (3). ©Ilaria Torre.

we prepared three sentences for each object, one describing the object in a positive way, one in a neutral way, and one in a negative way, for a total of 18 sentences. Thus, the avatar would describe the object positively when it moved the object to positions 1 or 2, neutrally when it moved the object to positions 3 or 4, and negatively when it moved the object to positions 5 or 6. The full list of sentences can be found in the Appendix.

In order to validate the valence of these 18 pre-scripted sentences, 15 English speakers, who self-identified as being at least fluent in the language, read each sentence, in random order, and rated whether the item in the sentence was described in a negative, neutral, or positive way, in a Likert scale going from -3 (valence = negative) to +3 (valence = positive). A cumulative linked mixed model with participants' rating as dependent variable, sentence valence and participants' level of English as predictors, and participant ID as random effect, was fitted to the data using the "ordinal" R package [19]. There was a significant main effect of valence: neutral sentences were 3.38 times more likely to be rated as neutral in valence than negative sentences ($z = 7.12$, $p < .001$), and positive sentences were 26.40 times more likely to be rated as positive in valence than negative sentences ($z = 14.14$, $p < .001$). There was no effect of English language fluency ($z = -1.27$, $p = .20$ for the native–fluent speaker comparison and $z = -1.69$, $p = .09$ for the near-native–fluent comparison), suggesting that participants rated the valence of the descriptions independently of their language abilities. The median rating of the positive descriptions was 2, the median of the neutral descriptions was 0, and the median of the negative descriptions was -2 (see Figure 3). Thus, we can be confident that fluent English speakers would interpret these sentences in a similar fashion.

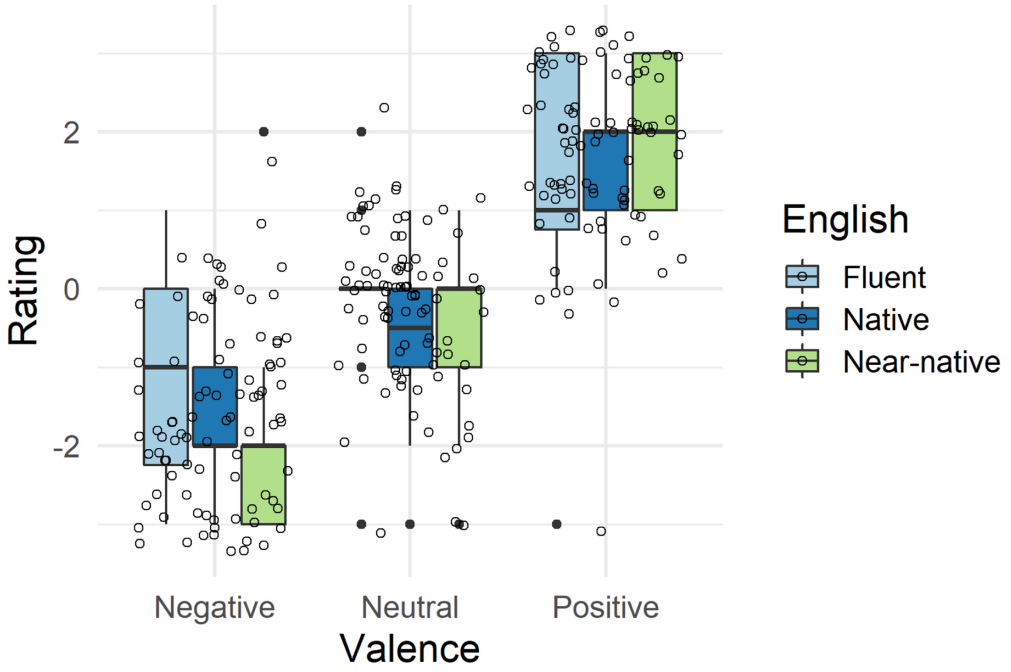


Fig. 3. Validation of the valence of the scripted utterances.

### 3.3 Motion Capture

The virtual character was created using state-of-the-art computer graphics technology for modelling, animating, and rendering. First, the model, comprising over 250 scans of a real actor's facial expressions, was created by a company called 3Lateral. These scans were then carefully combined into a controllable facial rig, which could then be driven by the motion capture. We then hired a male Irish actor to be recorded in our motion capture studio at Trinity College, Dublin. We used a 23-camera Vicon Vantage optical motion capture system for body motion capture and a Technoprops video-based head-mounted facial capture system (Figure 4). The actor was asked to read a set of pre-scripted sentences in neutral and smiling expressions (see Section 3.2). Audio was recorded using a wireless microphone attached to his face. The actor's facial movements were then retargeted onto the model, using Faceware Tech software for the facial movement and inverse kinematics for the movement of the head. Finally, advanced shaders (e.g., subsurface scattering for the skin) were used to create the highly realistic appearance in Autodesk Maya 2018 software. The final character is shown in Figure 5.

### 3.4 Video File Preparation

The audio recordings from the actor were processed using Audacity. First, a noise removal filter was applied to the recordings (with parameters: noise reduction 24dB, sensitivity 0dB, frequency smoothing 150 Hz, attack/decay 0.15 seconds); then the full audio file was segmented so as to obtain one file per utterance; finally, the individual sound files were amplitude-normalized. These files were then lip-synced to the individual, corresponding video files using Lightworks. Neutral sound files were lip-synced to neutral and smiling video files, and smiling sound files were lip-synced to smiling and neutral video files, to obtain the 4 desired experimental conditions: $V_sF_s$,


Fig. 4. Motion capture setup. Please note that the portrayed person is not the actor we recorded for this experiment. ©Ilaria Torre.

Fig. 5. Virtual character with a neutral (left) and smiling (right) facial expression.

Table 1. List of the Experimental Conditions for the Three Experiments

| | Code | Voice (V) | Face (F) |
|---|---|---|---|
| Experiment 1 | $V_sF_s$ | smiling | smiling |
| | $V_sF_n$ | smiling | neutral |
| | $V_nF_n$ | neutral | neutral |
| | $V_nF_s$ | neutral | smiling |
| Experiment 2 | $V_s$ | smiling | // |
| | $V_n$ | neutral | // |
| Experiment 3 | $F_s$ | // | smiling |
| | $F_n$ | // | neutral |

$V_sF_n$, $V_nF_n$, $V_nF_s$ (see Table 1). Note that in the matched condition we also lip-synced audio and video files originating from different recording takes. This meant that all stimuli, regardless of whether representing matched or mismatched conditions, were equally likely to present a slight asynchrony. We took this precaution in order to avoid confounds between synchrony and matched

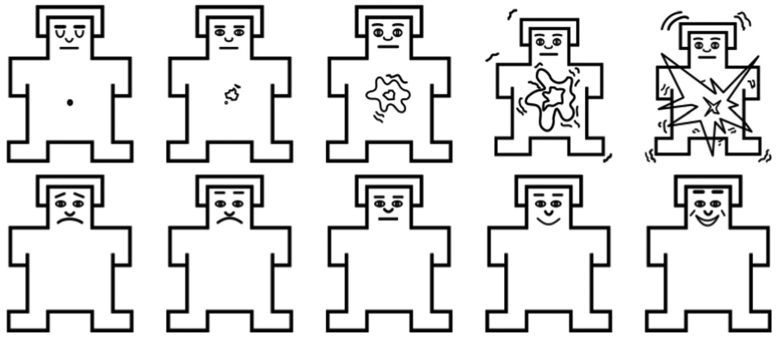

Fig. 6. The SAM visual scale used in the crowdsourced study: arousal (top row) and valence (bottom row).

stimuli, and asynchrony and mismatched stimuli. Nevertheless, we took maximum care to ensure that audio and video were as synchronous as possible.

## 4 ONLINE EXPERIMENT: SMILE VALIDATION

We validated the thus obtained audio-visual stimuli in two different ways. Firstly, we ran an online study on **Amazon Mechanical Turk** (**AMT**), where participants were asked to rate a subset of the stimuli in terms of valence and arousal, two orthogonal dimensions of emotions according to dimensional theories [64]. The stimuli were further validated during the main "lunar survival tasks" experiment, where participants were asked to rate the stimuli in terms of happiness (Section 5.3). Thus, we evaluated the stimuli both from a dimensional and a conceptual labelling perspective.

### 4.1 Procedure

We recruited 104 workers on AMT. They watched a total of 28 videos taken from our stimuli set, in random order. The videos featured two of the neutral utterances, so as to prevent participants from evaluating them based on their linguistic valence 3.2. These two utterances were presented in all 4 conditions ($V_nF_s$, $V_sF_s$, $V_nF_n$, and $V_sF_n$); we also showed 4 videos from the audio-only experiment ($V_n$ and $V_s$) and 4 from the video-only experiment ($F_n$ and $F_s$). The remaining 12 videos included a different artificial agent, and are therefore not discussed here. Participants watched one video at a time and were then asked to evaluate them using a visual scale, the SAM [13], shown in Figure 6. There was one scale for valence and one for arousal, each made of 5 points (where 1 corresponded to least aroused/lowest valence, and 5 to most aroused/highest valence). This evaluation study took approximately 10 minutes, and workers were compensated $3.2, which is higher than minimum wage in the USA, and higher than average Mechanical Turk pay rate [55].

### 4.2 Results

We excluded 6 participants who reported technical issues with the experiment. We then examined the time it took participants to evaluate each video; a rule of thumb to discover outliers is to filter out data points that are 3 standard deviations away from the mean. As reaction times cannot be negative, we therefore excluded individual trials for which participants took longer than 3 standard deviations away from the log of the mean of all participants. This resulted in the exclusion of 35 individual trials, for a total of 1,547 analysable trials. Participants (30 females, 66 males, 2 preferred not to say) were recruited from the United States; their self-reported English language fluency was: 93 native, 1 near-native, 4 fluent; they were aged 20–67 (mean = 38, sd = 11).

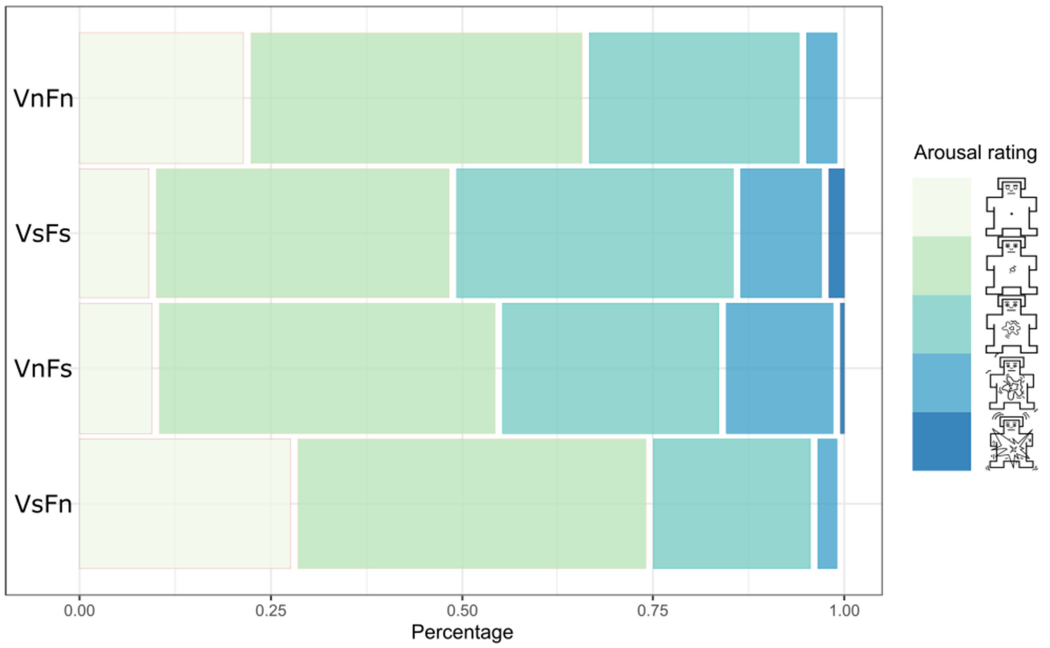


Fig. 7. Participants' arousal ratings in the audio-visual condition.

We fitted cumulative link mixed models to predict arousal and valence ratings, with smiling in the face and voice as predictors, and participant ID as random effect. We fitted these models separately for the audio-visual, audio only, and video only conditions, as these were between-subject conditions in our museum experiments. In the audio-visual condition, we found both face smiling and audio smiling to be significant predictors of arousal ratings (for face smiling: odds ratio = 6.24, 95% CI = [4.08, 9.56], $p < .001$; for voice smiling: odds ratio = 1.71, 95% CI = [1.14, 2.57], $p = .01$): as can be seen from Figure 7, the presence of smiling in the face and / or voice increased the arousal ratings. For valence, only face smiling was a significant predictor (for face smiling: odds ratio = 17.96, 95% CI = [11.29, 28.57], $p < .001$; for voice smiling: odds ratio = 0.97, 95% CI = [0.66, 1.43], $p = 0.89$): as can be seen in Figure 8, smiling in the face increased valence ratings.

For the audio only condition, smiling in the voice was a significant predictor for both arousal and valence (arousal: odds ratio = 3.26, 95% CI = [2.12, 5.01], $p < .001$; valence: odds ratio = 2.29, 95% CI = [1.43, 3.67], $p < .001$). Smiling in the voice increased ratings of arousal and valence in the audio-only condition.

Finally, for the video only condition, smiling in the face was a significant predictor for both arousal and valence (arousal: odds ratio = 4.06, 95% CI = [2.64, 6.24], $p < .001$; valence: odds ratio = 13.50, 95% CI = [8.19, 22.25], $p < .001$). Smiling in the face increases ratings of arousal and valence in the video-only condition.

## 5 MUSEUM EXPERIMENT 1: AUDIO-VISUAL MODALITY

The three lunar survival experiments were installed in a public setting in the Science Gallery, a museum adjacent to Trinity College Dublin, Ireland. The Science Gallery hosts regular exhibitions where artists and scientists present installations at the intersection of art and science. Visitors to the museum could go to a sign-up station manned by the museum staff; there they could fill

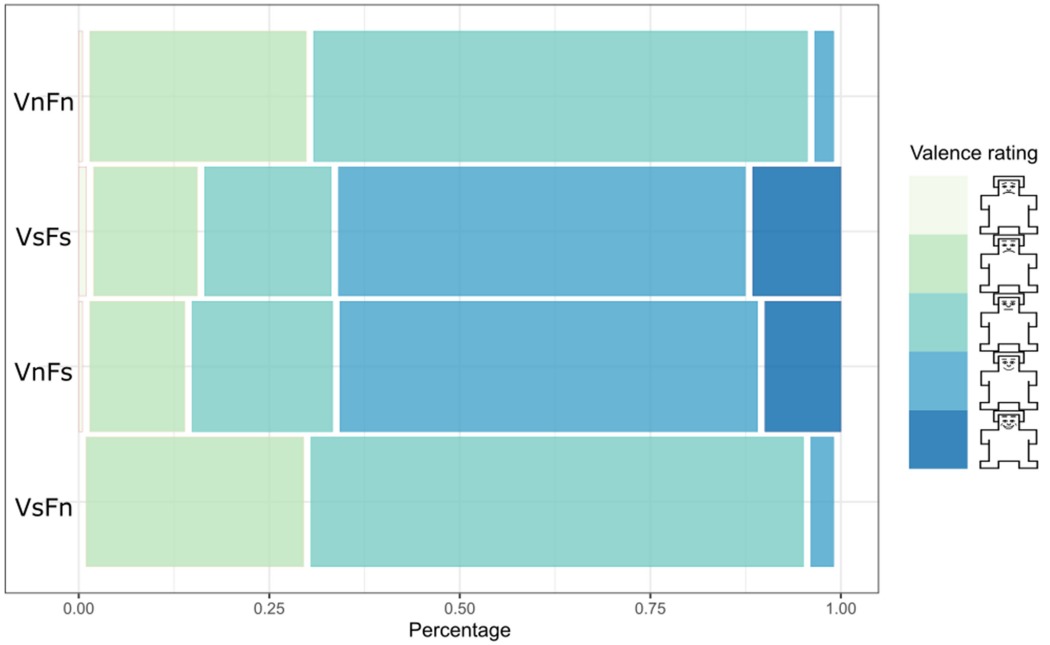


Fig. 8. Participants' valence ratings in the audio-visual condition.

out a demographic questionnaire on their gender, age, English language fluency, and country of origin, and be given a participant ID. They could then use this ID to play the "lunar survival task". Participants were seated in front of a computer monitor which had previously been calibrated to the following settings: brightness 200 cd/$m^2$, white point 6,500, display gamut 97% of sRGB; the character's video resolution was 1,280 × 720 and the monitor resolution was 1,920 × 1,080. This setup, as well as the data pre-processing pipeline (Section 5.2, applies to all three museum experiments.

### 5.1 Participants

Participants were visitors of the Science Gallery, who decided to interact with the experiment installation, and consented to their data being used for research purposes. While around 2,000 people interacted with the experiment over the course of its three-month residency, consent was obtained from 646 people for Experiment 1, 84 people for Experiment 2, and 105 people for Experiment 3.

### 5.2 Data Pre-Processing

As the experiments took place in an uncontrolled environment, without the constant presence of the experimenter to oversee the participants, it is possible that they were distracted during the experiment, or that this was otherwise disrupted. While we acknowledge that the resulting data is not as 'clean' as it would be in a standard laboratory setting, we believe that the inherent 'noise' in our data is a more accurate representation of people's behaviour, as this museum setting is much more natural and reflective of a real Human–Machine Interaction scenario than a laboratory setting. Still, we took some precautions to ensure that the data were suitably prepared for statistical analysis.

First of all, we looked at the free-text comments at the end of the questionnaire, and eliminated all participants whose comments made us believe that something had gone wrong with the experiment. For example, a few participants remarked that they had not been wearing the

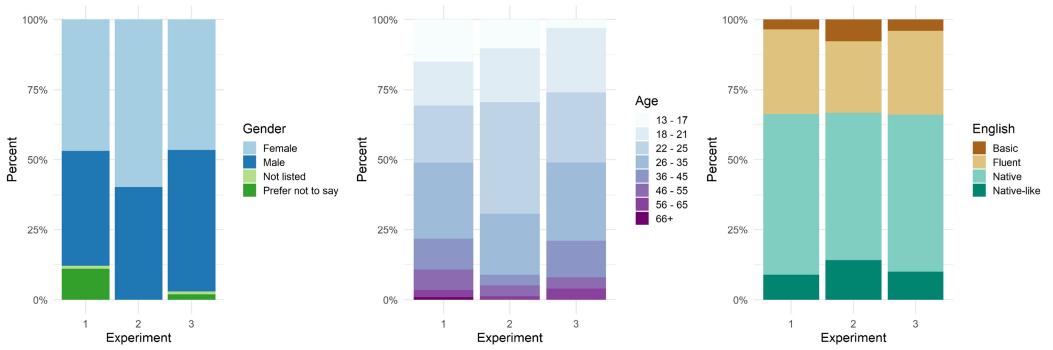


Fig. 9. Participants' demographics in the three experiments: gender (left), age range (middle), and English language fluency (right).

headphones during the game, so they could not hear the agent; or, other participants regretted not having read the task information at the beginning of the game. Secondly, we looked at the reaction times to see if certain people spent too long or too short a time to complete the game, as this could have been a signal that they were doing something else in the meantime, or that they were not paying enough attention. We calculated the log of all the reaction times and excluded all participants who were 3 SD away from the mean of these log times. Finally, we also eliminated the data points of participants who played the game more than once, although we retained the data from the first time they played.

This cleaning process resulted in a final participant sample of 597 people for Experiment 1, 78 people for Experiment 2, and 101 people for Experiment 3. A graphical summary of the demographics of the participants in the three experiments can be seen in Figure 9.

### 5.3 Procedure

Each participant was randomly assigned to one of the four experimental conditions (see Table 1). Participants first saw a screen with a description of the task and an empty field for their ID number. They were also asked to state whether this was their first time playing the game, and if they were wearing the provided headphones. This information was later used to filter data points (Section 5.2). After they clicked 'Next', they were taken to the first part of the game, where they saw a grid with six empty positions, and six icons representing the objects they had to rank (Figure 2, panel a). Hovering with the mouse over the icons would reveal the name of the objects, and these could be placed in the grid by dragging-and-dropping. After they made their first ranking, the virtual character videos were played, based on participants' ranking: the first video that was played was the one describing participant's object #1, which the virtual character moved to position #5; therefore, this was described in a negative way (see Section 3.2). The second video that was played was the one describing participant's object #2, and so on. While the videos were playing, the corresponding objects appeared in a second grid, in the position chosen by the virtual character (Figure 2, panel b). After all six videos were played, a third grid appeared, and the six objects were available to participants again for ranking (Figure 2, panel c). After participants made their final ranking and clicked 'Next', they were taken to a new page asking them to rate the virtual character on some characteristics using a Likert scale ranging from 1 (= not at all) to 5 (= very much). The characteristics were: realism, appeal, eeriness, trustworthiness, knowledgeableness, attractiveness, happiness, and intelligence. The questionnaire also contained a free-text box where participants could leave their comments on the experiment. The experiment ended when participants clicked 'Submit'.

### 5.4 Results

The data from the three experiments were analysed in the same way. Our main dependent variable was the Kendal-Tau distance[1] between the virtual character's ranking and the participants' final ranking. We interpreted this as an overall 'social influence score' of the agent: the smaller the number, the higher the influence.

We also calculated the distance between participants' initial ranking and an expert ranking: when the "lunar survival task" was first devised, 5 NASA crew members gave their ranking of the full list of 15 objects [as reported in 52]. These have been used as an expert baseline before [119]. The reasoning for including this distance was that people whose ratings were very similar to the expert ratings might have had more knowledge of the situation and been more certain of their initial ranking.

Since we found out that our assumption about participants not having any inherent object preference was wrong, as a post-hoc analysis we also calculated the changes in ranking order for objects #1, #2, #4, and #5. For object #1, this was a number ranging from 0 (no change) to 5 (moved 5 places down, as suggested by the virtual character); for object #2, the number ranged from 1 (moved 1 place up) to −4 (moved 4 places down, with −2 being the suggested change); for object #4, the number ranged from 3 (moved 3 places up, with 2 being the suggested change) to −2 (moved 2 places down); for object #5, it ranged from 4 (moved 4 places up, as suggested by the character) to −1 (moved 1 place down).

Finally, we computed the latent factors behind the perceived traits (realism, appeal, eeriness, trustworthiness, knowledgeableness, attractiveness, happiness, and intelligence). Factor analysis was performed using the 'psych' package [97]; post-hoc comparisons were estimated using the 'emmeans' package [67]. All analyses were performed in R version 3.5.1 [92].

*5.4.1 Does Emotional Expression Influence Behaviour?* We first looked at whether the virtual character's emotional expression influenced participants' final ranking. We fitted a linear model with the overall avatar social influence score as dependent variable and the character's emotional expression as predictor. The full model is reported in Table 2. There was no main effect of emotional expression. Adding distance from the expert ranking as an additional predictor did not improve model fit, and there was no interaction.

We then added the perceived traits to the regression model to see if perceiving the character in a certain way increased its influence. Intelligence, knowledgeableness, happiness, and trustworthiness (approaching significance at the 5% level) were found as main effects. Specifically, people trusted the character more if they perceived it to be intelligent, knowledgeable, and trustworthy, but they trusted it less if they perceived it to be happy.

We then added the participants' demographic information to the model to account for any individual differences in people's behaviour. There was no effect of gender or age, but there was a main effect of English language fluency. Specifically, people who were native English speakers accepted the character's recommendations more than near-native speakers, who in turn accepted them more than fluent speakers, who accepted them more than basic speakers. As this task requires being able to understand sentences in the English language, it is possible that people who might have not understood everything the agent said were less inclined to follow its advice. After all, it is well established in Sociolinguistics that people trust speakers of their own language more than others, and this preference emerges from a very young age [83, 88]. Since utterance validation suggested that people who are at least fluent in English should be able to understand their

[1]This distance is a representation of the relationship between two vectors of ranked data; in this case, the vectors contain the ranking made by the virtual agent and the final ranking made by the participant.

Table 2. Full Regression Model for Experiment 1, Showing the Model Predictors, Including Significant Interactions and Intercept, $\beta$ Estimates and 95% Confidence Intervals in Parentheses

| | *Dependent variable:* |
|---|---|
| | Social influence score |
| Intercept | $10.390^{***}$ (8.054, 12.726) |
| Expression: $V_sF_s$ | −0.451 (−2.307, 1.405) |
| Expression: $V_nF_s$ | 0.011 (−1.711, 1.734) |
| Expression: $V_sF_n$ | −0.558 (−2.240, 1.125) |
| Distance from expert | 0.050 (−0.123, 0.224) |
| Realism | 0.139 (−0.102, 0.380) |
| Appeal | 0.007 (−0.280, 0.295) |
| Eerie | 0.144 (−0.073, 0.362) |
| Attractiveness | −0.133 (−0.422, 0.156) |
| Knowledgeableness | $-0.627^{***}$ (−0.972, −0.281) |
| Intelligence | $-0.621^{***}$ (−0.966, −0.277) |
| Trustworthiness | $-0.269^{*}$ (−0.544, 0.007) |
| Happiness | $0.288^{**}$ (0.036, 0.539) |
| English: Fluent | $-2.217^{***}$ (−3.555, −0.879) |
| English: Native | $-2.638^{***}$ (−3.949, −1.327) |
| English: Native-like | $-2.488^{***}$ (−3.959, −1.016) |
| Age: 18–21 | −0.269 (−1.105, 0.568) |
| Age: 22–25 | −0.407 (−1.199, 0.386) |
| Age: 26–35 | 0.056 (−0.707, 0.819) |
| Age: 36–45 | −0.567 (−1.512, 0.378) |
| Age: 46–55 | 0.318 (−0.750, 1.385) |
| Age: 56–65 | −0.523 (−2.079, 1.033) |
| Age: 66+ | −2.100 (−4.896, 0.696) |
| Gender: Male | −0.146 (−0.623, 0.331) |
| Gender: Not listed | 0.076 (−2.185, 2.338) |
| Gender: Prefer not to say | 1.051 (−0.983, 3.084) |
| Expression: $V_sF_s$ X distance from expert | −0.024 (−0.288, 0.239) |
| Expression: $V_nF_s$ X distance from expert | −0.094 (−0.347, 0.160) |
| Expression: $V_sF_n$ X distance from expert | 0.044 (−0.205, 0.292) |
| $R^2$ | 0.230 |
| Adjusted $R^2$ | 0.187 |

*Note:* $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

linguistic content (Section 3.2), we ran a follow-up analysis excluding people who self-reported as basic speakers of English. We found the same effects for emotional expression, distance from expert, intelligence, knowledgeableness, and trustworthiness. The effect of happiness was no longer significant, however ($F(1) = 1.09$, $p = .30$).

*5.4.2 Are There Preferences Towards Certain Objects?* Since we did not find a main effect of avatar emotional expression on social influence overall, we zoomed in on the possible effect of different objects and relative positions. Although our starting assumption was that each object would be equally likely to be ranked as most important (see Section 3.1), we ran a series of post-hoc analyses to see whether people had strong preferences towards a certain object. Specifically, we looked

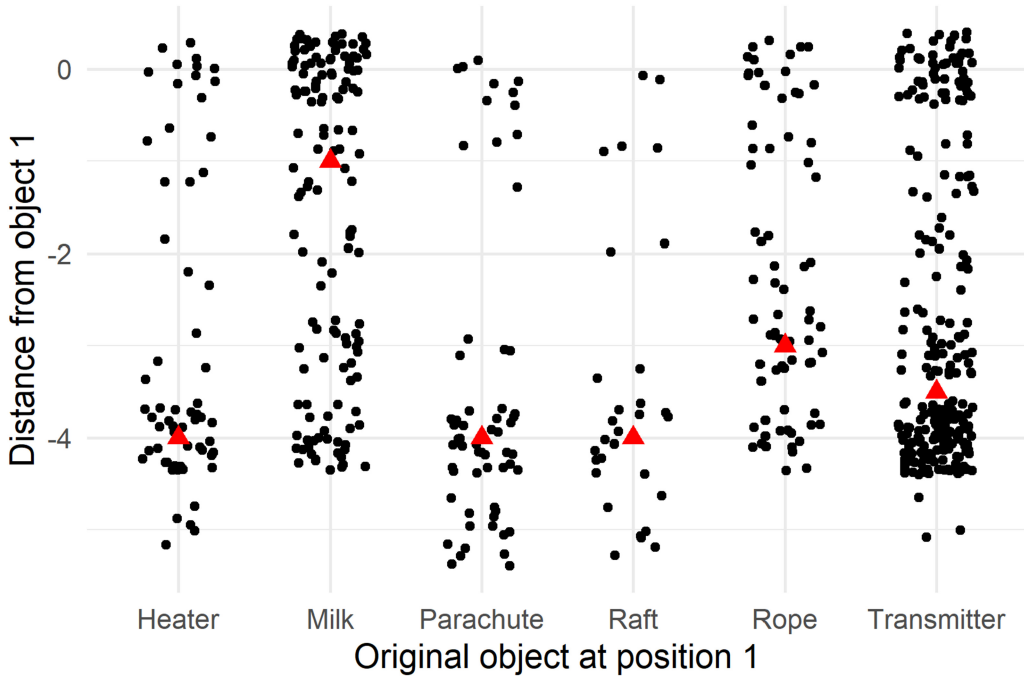


Fig. 10. Objects originally placed at position #1 by participants, shown in relation to how often they were demoted according to the agent's suggestion. Data on the *x*- and *y*-axes were jittered to improve display. The red triangle represents the median.

at whether people were more likely to change the position of only certain objects. We performed a Kruskal–Wallis rank sum test on each object position (excluding positions #3 and #6, which were not changed by the virtual character) and, when this test was significant, we performed post-hoc comparisons using Dunn's test of multiple comparisons, with Benjamini-Hochberg adjustment.

The Kruskal–Wallis rank sum test was significant for objects initially placed at position #1 ($\chi^2(5) = 84.25$, $p < .001$, Figure 10). Post-hoc comparisons showed that the milk was demoted less than the heater ($Z = 5.27$, $p < .001$), the parachute ($Z = 7.48$, $p < .001$), the raft ($Z = 5.83$, $p < .001$), and the transmitter ($Z = 5.55$, $p < .001$); the rope was demoted less than the heater ($Z = 3.00$, $p < .001$), the parachute ($Z = 4.98$, $p < .001$), the raft ($Z = 4.06$, $p < .001$), and the transmitter ($Z = 2.30$, $p = .03$); finally, the transmitter was demoted less than the parachute ($Z = -3.89$, $p < .001$) and less than the raft ($Z = -2.95$, $p < .01$).

The Kruskal–Wallis rank sum test was also significant for objects initially placed at position #2 ($\chi^2(5) = 28.74$, $p < .001$, Figure 11). Post-hoc comparisons showed that the milk was demoted less than the heater ($Z = 4.11$, $p < .001$), the raft ($Z = 4.33$, $p < .001$), and the transmitter ($Z = 4.07$, $p < .001$); the rope was demoted less than the heater ($Z = 2.62$, $p = .03$), the raft ($Z = 3.18$, $p < .01$), and the transmitter ($Z = 2.30$, $p = .03$); finally, the rope was demoted less than the transmitter ($Z = 2.53$, $p = .03$).

The Kruskal–Wallis rank sum test for objects initially placed at position #4 approached significance at the 5% level ($\chi^2(5) = 9.66$, $p = .09$). Post-hoc comparisons failed to find any significant difference between objects.

The Kruskal–Wallis rank sum test was significant for objects initially placed at position #5 ($\chi^2(5) = 26.69$, $p < .001$, Figure 12). Post-hoc comparisons showed that the raft was promoted

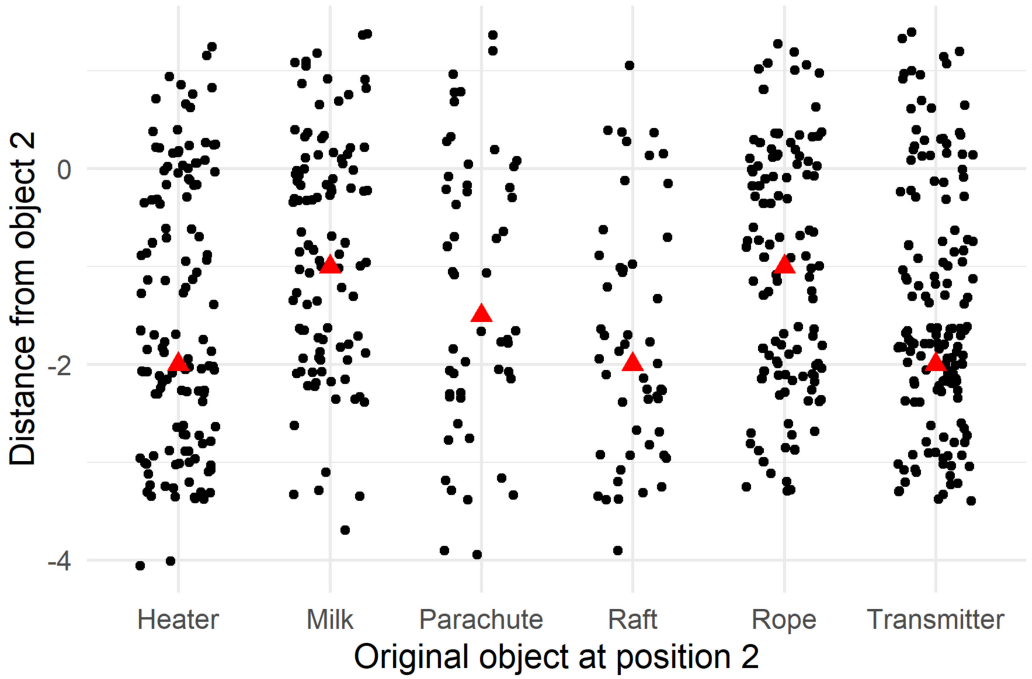


Fig. 11. Objects originally placed at position #2 by participants, shown in relation to how often they were demoted according to the agent's suggestion. Data on the *x*- and *y*-axes were jittered to improve display. The red triangle represents the median.

less than the heater ($Z = -2.72$, $p = .02$), the parachute ($Z = -3.64$, $p < .01$), the rope ($Z = -2.36$, $p = .03$), and the transmitter ($Z = -4.46$, $p < .001$); the transmitter was also promoted more than the heater ($Z = 2.57$, $p = .03$), the parachute ($Z = 2.37$, $p = .04$), and the rope ($Z = 2.73$, $p = .03$).

*5.4.3 Does Emotional Expression Have a Different Effect Based on the Item's Original Position?* Given that people seemed more reluctant to move objects that they placed in certain positions, we also looked at whether the emotional expression influenced the rankings for each initial object position separately. Linear regression showed a main effect of emotional expression for objects initially placed at position #1 (Table 3); post-hoc comparisons using the Tukey HSD method showed that the $V_nF_s$ condition had more influence than the $V_sF_s$ condition. There was also a main effect of distance from expert ranking: as can be seen from Figure 13, the more people deviate from expert rankings, the less they accept the avatar's rankings.

We also found an interaction between emotional expression and distance from expert rankings for objects initially placed at position #4. As can be seen from Figure 14, the more people differ from the expert ranking, the less they follow the recommendations of the $V_sF_s$ avatar and the more they follow the $V_nF_s$ avatar.

*5.4.4 Factor Analysis.* The eight trait ratings in the perceptual questionnaire were factor-analysed using principal component analysis with Varimax (orthogonal) rotation. The analysis yielded two factors explaining a total of 73.3% of the variance for the entire set of traits. The analysis showed that the perceived traits cluster around two groups: knowledgeableness, intelligence, and trustworthiness belong to one group, and appeal, attractiveness, realism, eeriness, and

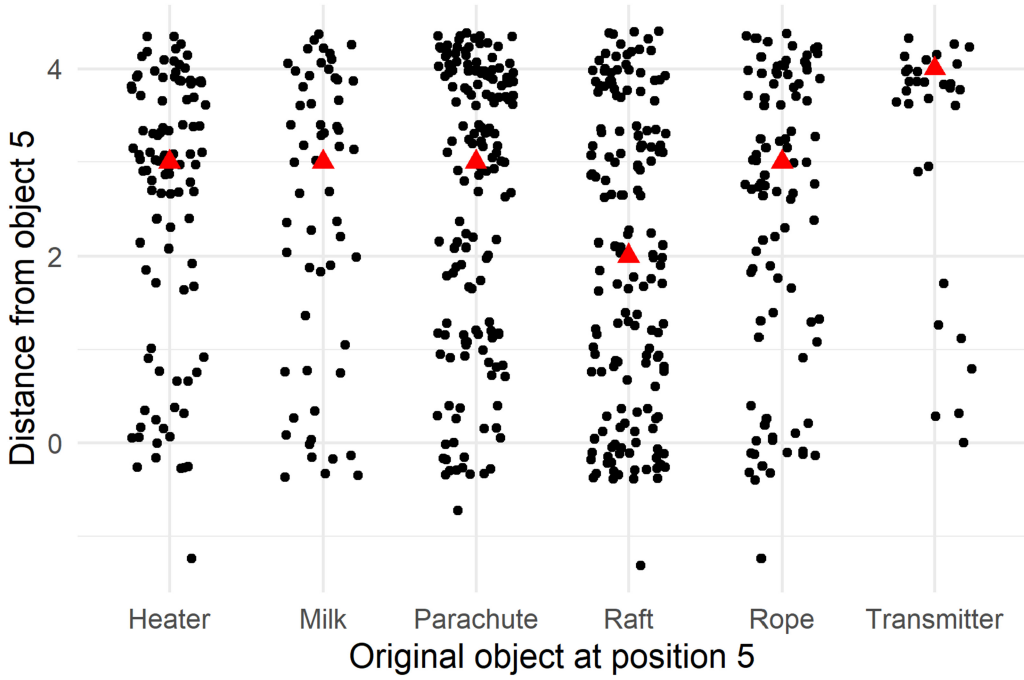


Fig. 12. Objects originally placed at position #5 by participants, shown in relation to how often they were promoted according to the agent's suggestion. Data on the *x*- and *y*-axes were jittered to improve display. The red triangle represents the median.

Table 3. Regression Models for the Distance from Objects Originally Placed at Positions #1, #2, #4, and #5 in Experiment 1

| Dependent variable | Distance from #1 | Distance from #2 | Distance from #4 | Distance from #5 |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Intercept | −3.107*** (−3.827, −2.386) | −1.340*** (−1.870, −0.811) | 1.884*** (1.346, 2.422) | 2.347*** (1.708, 2.986) |
| $V_sF_s$ | 0.244 (−0.871, 1.358) | 0.147 (−0.673, 0.967) | 0.532 (−0.300, 1.364) | −0.084 (−1.072, 0.904) |
| $V_nF_s$ | −0.259** (−1.293, 0.775) | 0.705 (−0.055, 1.465) | −0.632 (−1.404, 0.139) | 0.438 (−0.478, 1.355) |
| $V_sF_n$ | 0.421 (−0.589, 1.432) | 0.086 (−0.657, 0.829) | −0.308 (−1.063, 0.446) | 0.051 (−0.844, 0.947) |
| Distance-expert | 0.111** (0.006, 0.216) | −0.013 (−0.091, 0.064) | −0.047 (−0.125, 0.032) | −0.010 (−0.103, 0.083) |
| $V_sF_s$ : distance-expert | −0.025 (−0.184, 0.135) | −0.009 (−0.126, 0.108) | −0.095 (−0.213, 0.024) | 0.032 (−0.109, 0.173) |
| $V_nF_s$ : distance-expert | −0.023 (−0.177, 0.131) | −0.075 (−0.188, 0.038) | 0.094** (−0.021, 0.209) | −0.027 (−0.164, 0.109) |
| $V_sF_n$ : distance-expert | −0.087 (−0.238, 0.064) | 0.024 (−0.087, 0.135) | 0.042 (−0.070, 0.155) | 0.026 (−0.108, 0.160) |
| $R^2$ | 0.028 | 0.015 | 0.021 | 0.006 |
| Adjusted $R^2$ | 0.017 | 0.003 | 0.010 | −0.006 |

*Note:* *p < 0.1; **p < 0.05; ***p < 0.01.
It shows the model predictors, including interactions and intercept, $\beta$ estimates and 95% confidence intervals in parentheses.

happiness to another. A Kruskal–Wallis rank sum test then showed that the virtual character's emotional expression influenced these ratings ($\chi^2(3) = 7.86$, $p = .05$): in particular, the $V_nF_s$ condition scored higher than $V_sF_s$ in terms of knowledgeableness, intelligence, and trustworthiness (Dunn test, $Z = 2.69$, $p = .04$).

*5.4.5 Perceptual Survey.* Looking at individual trait comparisons, we found that people rated the $V_nF_s$ condition as more appealing than $V_sF_s$ (Kruskal–Wallis: $\chi^2(3) = 8.01$, $p = .05$; Dunn test:

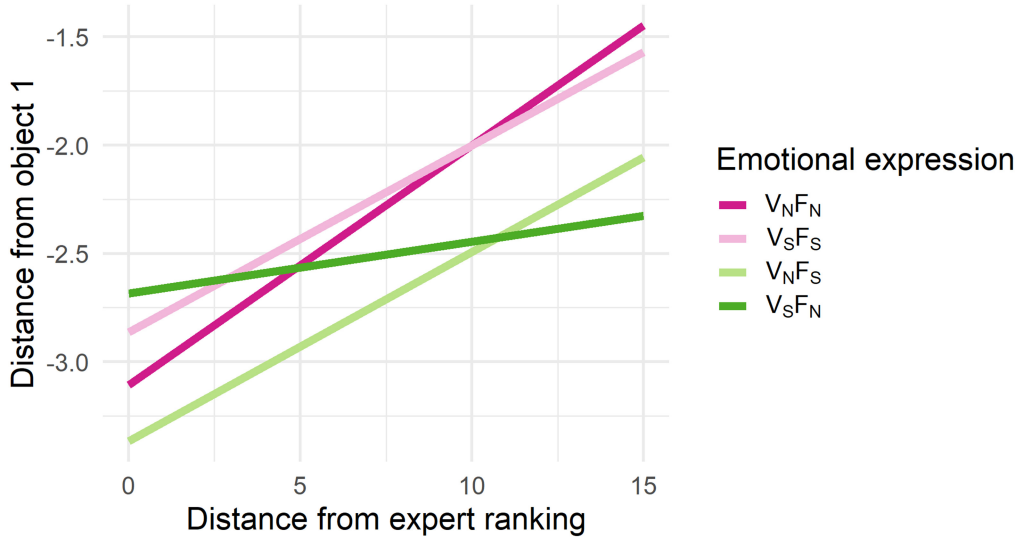


Fig. 13. Main effect of distance from expert ranking: people who placed objects at position #1 similarly to experts were more likely to accept the avatar's suggested changes, in all four emotional expression conditions.

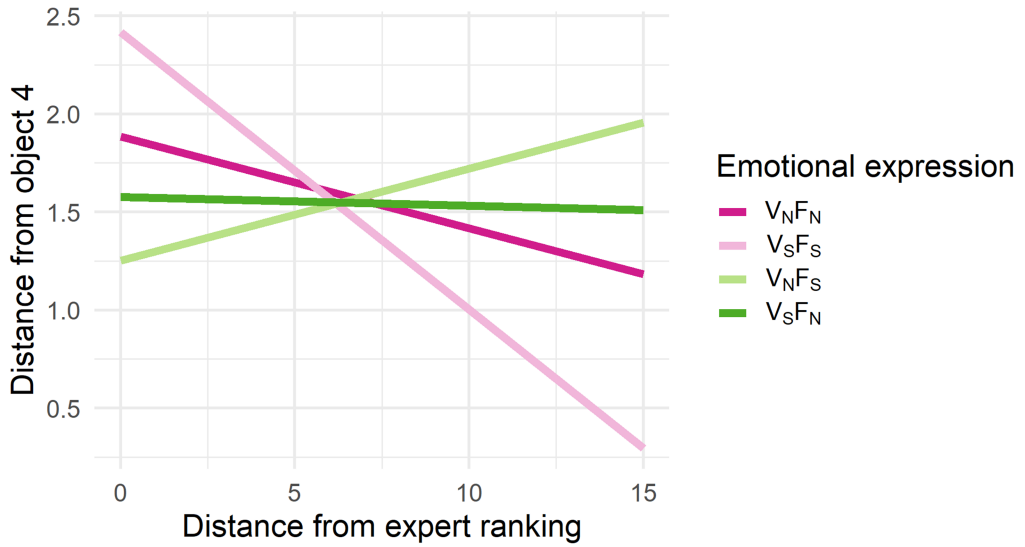


Fig. 14. Interaction between emotional expression and distance from expert ranking: people who placed objects at position #4 differently from experts were more likely to accept the avatar's suggested changes in the $V_nF_s$ condition, and less likely to accept them in the $V_sF_s$ condition.

$Z = 2.69$, $p = .04$); people rated $V_sF_s$ and $V_nF_s$ as happier than $V_nF_n$ and $V_sF_n$ (Kruskal–Wallis: $\chi^2(3) = 35.87$, $p < .001$; Dunn tests: $V_sF_s$ - $V_nF_n$ $Z = 4.92$, $p < .001$; $V_sF_s$ - $V_sF_n$: $Z = 4.23$, $p < .001$; $V_nF_s$ - $V_nF_s$: $Z = 4.08$, $p < .001$; $V_nF_s$ - $V_sF_n$: $Z = 3.42$, $p < .001$). They also rated $V_nF_s$ as more knowledgeable than $V_sF_s$ (Kruskal–Wallis: $\chi^2(3) = 7.90$, $p = .05$; Dunn test: $Z = 2.74$, $p = .04$). Finally, they rated $V_nF_s$ as more trustworthy than $V_sF_s$ (Kruskal–Wallis: $\chi^2(3) = 11.08$,

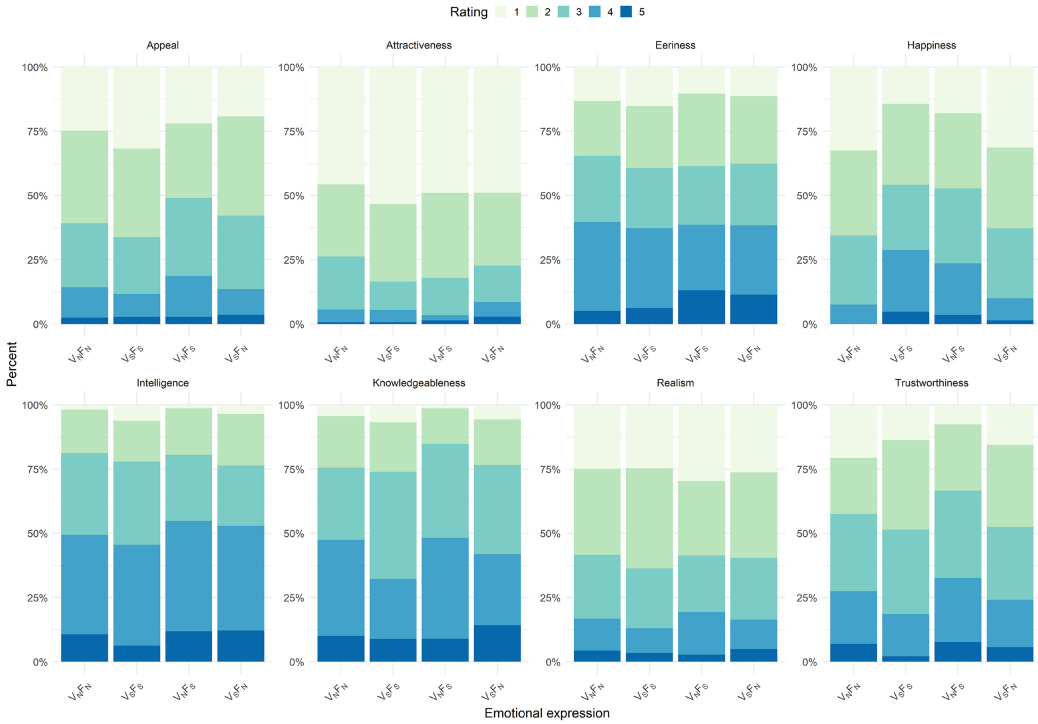


Fig. 15. Participants' ratings of the virtual character in terms of appeal, attractiveness, eeriness, happiness, intelligence, knowledgeableness, realism, and trustworthiness.

$p = .01$; Dunn test: $Z = 3.12, p = .01$), $V_sF_n$ ($Z = 2.54, p = .03$), and $V_nF_n$ (approaching significance at the 5% level: $Z = 2.17, p = .06$). In general, it seems that the $V_nF_s$ condition scored higher on the first latent factor—encompassing intelligence, knowledgeableness, and trustworthiness (Anova: $F(3) = 2.83, p = .04$, Tukey HSD: $p = .03$). The ratings for all the traits can be seen in Figure 15.

## 6 MUSEUM EXPERIMENT 2: AUDIO MODALITY

### 6.1 Procedure

In this experiment, participants played the same "lunar survival task", but without seeing the virtual character's face, only hearing its voice. For this reason, the experimental conditions, to which participants were randomly assigned, were only 2: $V_s$ and $V_n$. Otherwise, the procedure was exactly the same as Experiment 1, with the difference that the box where the virtual character's videos were played in Experiment 1 (Figure 2), was black in Experiment 2.

### 6.2 Results

*6.2.1 Does Emotional Expression Influence Behaviour?* We fitted a linear model with the overall participants' trusting score as the dependent variable and the character's emotional expression and distance from the experts as predictors. We found no main effects and no interaction. Adding the perceived traits to the model, we found that perceiving the voice as trustworthy and intelligent decreased the overall distance from the character's ranking—and therefore, increased the agent's social influence. Finally, we added participants' demographic information as predictors in the model, and found a main effect of age and a marginally significant interaction between gender

Table 4. Full Regression Model for Experiment 2, Showing the Model Predictors, Including Significant Interactions and Intercept, $\beta$ Estimates and 95% Confidence Intervals in Parentheses

| Dependent variable | Social influence score |
|---|---|
| Intercept | $10.118^{***}$ (6.669, 13.566) |
| $V_s$ | 0.580 (−0.469, 1.628) |
| Distance from expert | −0.0004 (−0.204, 0.203) |
| Attractiveness | −0.246 (−0.889, 0.397) |
| Appeal | −0.202 (−0.889, 0.486) |
| Realism | −0.039 (−0.595, 0.518) |
| Trustworthiness | $-0.582^{***}$ (−1.155, −0.009) |
| Happiness | 0.118 (−0.604, 0.839) |
| Intelligence | $-0.594^{**}$ (−1.125, −0.063) |
| Knowledgeableness | −0.086 (−0.709, 0.538) |
| Eerie | −0.291 (−0.826, 0.243) |
| Male gender | $-1.046^{*}$ (−2.135, 0.042) |
| Age 18–21 | −0.833 (−2.624, 0.957) |
| Age 22–25 | $-1.459^{*}$ (−3.058, 0.140) |
| Age 26–35 | 0.987 (−0.793, 2.766) |
| Age 36–45 | $-3.811^{***}$ (−6.603, −1.018) |
| Age 46–55 | $-3.056^{**}$ (−5.789, −0.324) |
| Age 56–65 | −1.888 (−6.204, 2.429) |
| Trustworthiness : Male gender | $-0.316^{*}$ (−0.634, 0.001) |
| $R^2$ | 0.444 |
| Adjusted $R^2$ | 0.368 |

*Note:* $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

and perceived trustworthiness. Specifically, people aged 26–35 had an overall higher distance from the character's rankings, i.e., they were influenced less; and men who perceived the voice as trustworthy had a smaller overall distance from the character's ranking, i.e., they were influenced more. The full model is shown in Table 4.

*6.2.2 Are There Preferences Towards Certain Objects?* Again, we looked at whether people were more reluctant to change the position of specific objects in the audio only experiment. We performed a Kruskal–Wallis rank sum test on each object position (excluding positions #3 and #6, which were not changed by the virtual character) and, when this test was statistically significant, we performed post-hoc comparisons using Dunn's test for multiple comparisons, with Benjamini-Hochberg adjustment. The Kruskal–Wallis rank sum test was significant for objects originally placed at position #1 ($\chi^2(4) = 10.75$, $p = .03$); post-hoc comparisons showed that people tended to change the position of the parachute more than the position of the milk (approaching significance at the 5% level: $Z = 2.71$, $p = .07$) and more than the position of the rope (approaching significance at the 5% level: $Z = 2.42$, $p = .08$). There was no difference in demoting different objects from position #2 ($\chi^2(5) = 4.32$, $p = .50$). For position #4, the Kruskal–Wallis test was marginally significant at the 5% level ($\chi^2(5) = 10.3$, $p = .07$), but post-hoc analyses failed to find any significantly different pairs, after correction for multiple comparisons. Finally, there were no differences in object changes when the objects had originally been placed at position #5 ($\chi^2(5) = 6.77$, $p = .24$).

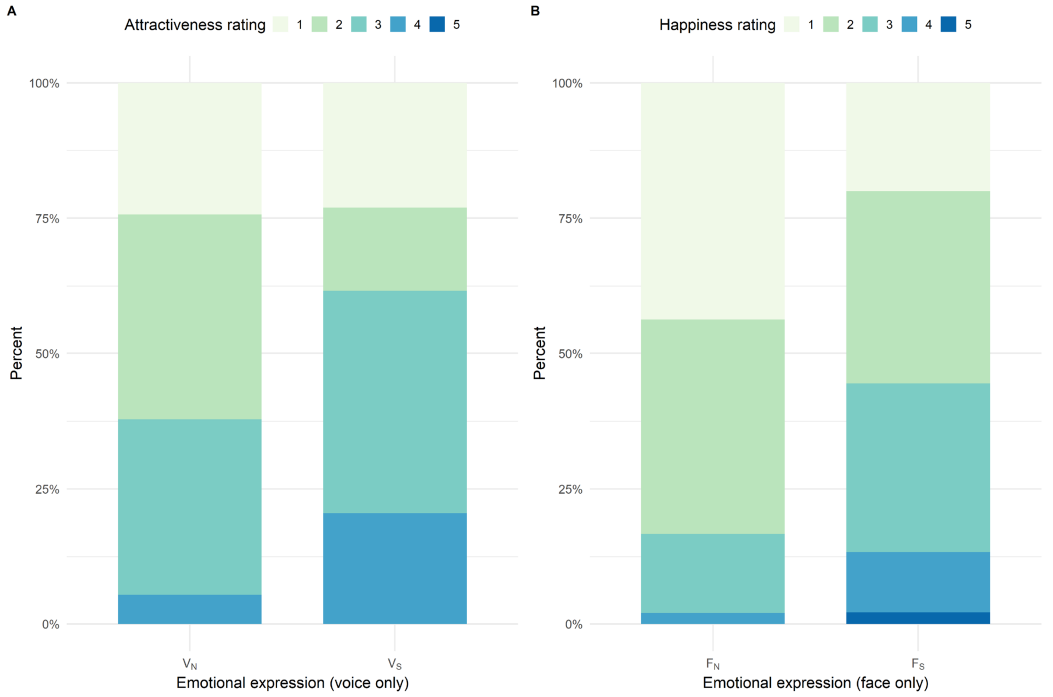


Fig. 16. Participants' ratings of the voice-only virtual character in terms of attractiveness (panel A) and of the face-only virtual character in terms of happiness (panel B).

*6.2.3 Does Emotional Expression Have a Different Effect Based on the Item's Original Position?* We also looked at whether the character's emotional expression influenced the rankings for each initial position separately. There was no significant effect for objects initially placed at positions #1, #2, and #4, but there was a main effect for objects initially placed at position #5 ($F(1) = 5.31$, $p = .02$). In particular, people trusted the $V_n$ condition more ($\beta = -0.78$). There was no main effect of distance from the expert ranking, and no interaction.

*6.2.4 Factor Analysis.* The eight trait ratings in the final questionnaire were factor-analysed using principal component analysis with Varimax (orthogonal) rotation. The analysis yielded two factors explaining a total of 72.1% of the variance for the entire set of traits. The analysis showed that the perceived traits cluster around two groups: attractiveness, appeal, realism, trustworthiness, and happiness belong to one group, and knowledgeableness, intelligence, and eeriness, to another. A Kruskal–Wallis rank sum test showed that the virtual character's emotional expression did not influence these two factor scores ($\chi^2(1) = 0.87$, $p = .35$).

*6.2.5 Perceptual Survey.* The virtual character's emotional expression marginally affected participants' individual ratings, but only in terms of attractiveness ($\chi^2(1) = 3.29$, $p = .07$), with the $V_s$ condition being rated as more attractive than the $V_n$ condition (Figure 16, panel A).

## 7 MUSEUM EXPERIMENT 3: VIDEO MODALITY

### 7.1 Procedure

In this experiment, participants played the same "lunar survival task", but without hearing the virtual character's voice, only seeing its face. For this reason, the experimental conditions, to

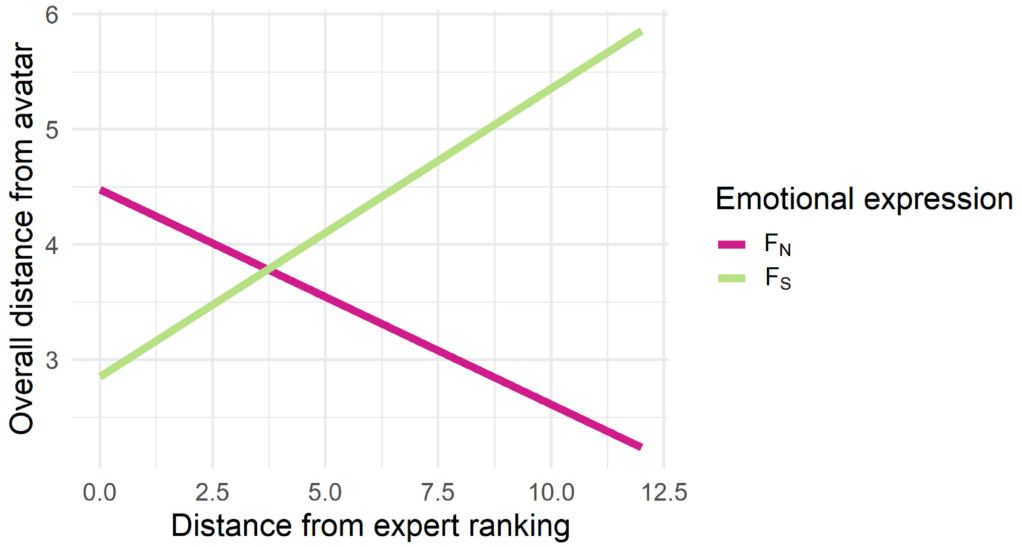


Fig. 17. Interaction between emotional expression and distance from expert ranking: people who rank their objects more differently from the experts are more likely to be influenced overall by the neutral agent than the smiling one.

which participants were randomly assigned, were only 2: $F_s$ and $F_n$. The avatar's sentences were displayed in text on the screen, as if they were subtitles. Otherwise, the procedure was exactly the same as Experiment 1.

### 7.2 Results

*7.2.1 Does Emotional Expression Influence Behaviour?* We fitted a linear model with the overall social influence score as dependent variable and the character's emotional expression and the distance from expert as predictors, and we found a main effect of emotional expression, with participants accepting the recommendations from $F_n$ more than $F_s$. Post-hoc analyses revealed that this effect was strongest for objects originally placed at position #1 ($F(1) = 5.00$, $p = .03$, $\beta = 0.76$). We also found an interaction between avatar emotional expression and distance from experts. As can be seen from Figure 17, people who deviate more from the expert rankings are more convinced by the neutral avatar than the smiling avatar.

Adding the perceived traits to the model, we found main effects of perceived trustworthiness, knowledgeableness, and eeriness. Specifically, perceiving the character as trustworthy and knowledgeable decreased the overall distance from its rankings, i.e., increased its social influence, while perceiving it as eerie tended to decrease its social influence. Finally, adding participants' demographic information to the model did not improve model fit, suggesting that individual differences did not play a role in the overall behaviour towards the face-only agent. The full regression coefficients can be seen in Table 5.

*7.2.2 Are There Preferences Towards Certain Objects?* Again, we looked at whether people were more likely to change the position only for some of the objects. We ran a Kruskal–Wallis rank sum test on each object position (excluding positions #3 and #6, which were not changed by the virtual character) and, when this test was statistically significant, we performed post-hoc comparisons using Dunn's test for multiple comparisons, with Benjamini-Hochberg adjustment. The

Table 5. Full Regression Model for Experiment 3, Showing the Model Predictors, Including Significant Interactions and Intercept, $\beta$ Estimates and 95% Confidence Intervals in Parentheses

| | *Dependent variable:* |
|---|---|
| | Social influence score |
| $F_s$ | $-1.624^{**}$ ($-4.183$, 0.935) |
| Distance from expert | $-0.186$ ($-0.424$, 0.051) |
| Attractiveness | 0.392 ($-0.313$, 1.098) |
| Appeal | 0.506 ($-0.168$, 1.180) |
| Realism | 0.072 ($-0.422$, 0.567) |
| Trustworthiness | $-0.116^{**}$ ($-0.627$, 0.395) |
| Happiness | 0.052 ($-0.512$, 0.617) |
| Knowledgeableness | $-1.008^{***}$ ($-1.523$, $-0.493$) |
| Intelligence | $-0.656^{*}$ ($-1.329$, 0.016) |
| Eerie | $0.351^{*}$ ($-0.041$, 0.744) |
| Male gender | 0.793 ($-0.225$, 1.811) |
| Age 18–21 | 0.380 ($-2.425$, 3.185) |
| Age 22–25 | 0.579 ($-2.224$, 3.382) |
| Age 26–35 | $-0.481$ ($-3.268$, 2.306) |
| Age 36–45 | $-1.023$ ($-3.931$, 1.885) |
| Age 46–55 | $-0.044$ ($-3.557$, 3.469) |
| Age 56–65 | 0.610 ($-2.871$, 4.091) |
| $F_s$ : Distance from expert | $0.437^{**}$ (0.062, 0.811) |
| Intercept | $4.987^{**}$ (0.557, 9.417) |
| $R^2$ | 0.378 |
| Adjusted $R^2$ | 0.224 |

*Note:* $^{*}$p < 0.1; $^{**}$p < 0.05; $^{***}$p < 0.01.

Kruskal–Wallis rank sum test was significant for objects originally placed at position #1 ($\chi^2(5) = 15.62$, $p < .01$); post-hoc comparisons showed that people who had placed the milk at position #5 demoted it less than the heater ($Z = 3.22$, $p < .01$), the parachute ($Z = 3.45$, $p < .01$), and the transmitter ($Z = 2.87$, $p = .02$). The test approached significance at the 5% level for objects originally placed at position #2 ($\chi^2(5) = 10.49$, $p = .06$); post-hoc comparisons showed that people who had placed the milk at position #2 demoted it less than the heater ($Z = 3.06$, $p = .03$). There was no difference in promoting objects originally placed at position #4 ($\chi^2(5) = 2.53$, $p = .77$) and #5 ($\chi^2(5) = 2.09$, $p = .84$).

*7.2.3 Factor Analysis.* The eight trait ratings in the final questionnaire were factor-analysed using principal component analysis with Varimax (orthogonal) rotation. The analysis yielded two factors explaining a total of 52.6% of the variance for the entire set of traits. The analysis showed that the perceived traits cluster around two groups: knowledgeableness, intelligence, and trustworthiness belong to one group, and appeal, realism, attractiveness, happiness, and eeriness, to another. A Kruskal–Wallis rank sum test showed that the virtual character's emotional expression did not influence these two factor scores ($\chi^2(1) = 0.61$, $p = .43$).

*7.2.4 Perceptual Survey.* The virtual character's emotional expression affected participants' individual ratings only in terms of happiness ($\chi^2(1) = 11.60$, $p < .001$), with $F_s$ being rated as happier than $F_n$ (Figure 16, panel B).

## 8 GENERAL DISCUSSION

### 8.1 Influence of Emotional Expressivity on Behaviour

We looked at how a virtual character's emotional expression, manipulated in the voice and face, affected its social influence in a "lunar survival task" [52]. The survival task has been used successfully in the past, showing that generally people are willing to accept recommendations from artificial agents of various kinds [e.g., 18, 87, 112, 119].

We found that our manipulation only partially influenced participants' behaviour: in the first museum experiment, where people interacted with a virtual character that had both a face and a voice, we found that overall people did not change their behaviour in the different avatar emotional expression conditions. After a follow-up exploration of the data, we found that participants tended to accept the agent's suggestions in the $V_nF_s$ condition more, but only in some cases, as this effect was mediated by the specific position where participants had initially placed the objects.

In the second experiment, where people only heard the agent's voice, again we did not find a main effect of emotional expressivity. People tended to accept the agent's suggestions in the $V_n$ condition more, but again this was mediated by the initial object location.

Finally, in Experiment 3, where people only saw the agent's face, people overall accepted the suggestions in the $F_n$ condition more, regardless of the object's initial position.

Thus, we find limited evidence that people followed the advice of an avatar displaying an incongruent, mismatched emotional expression when both channels were available, and a neutral expression when only one channel was available.

These combined results suggest that the character's smiling voice contains some features that were particularly disliked in the context of the game. This is apparent in the case of Experiment 3, where people overall accepted the recommendation from the neutral agent more. As for Experiment 1, if the visual information had overridden auditory information in the perception of the agent, people would have equally accepted its recommendations in the $V_sF_s$ and $V_nF_s$ conditions, but this was not the case—confirming what we found in our initial analyses [114]. Based on previous research on the cross-modal integration in emotion processing [43, 54, 86], it is possible that, in the case of multi-modal emotional expression, people trusted a character that was expressing a positive emotion (valence, in this case smiling in the face) and low activation (neutral voice); on the other hand, in the case of an emotion expressed through a single modality, be it voice or face, people trusted a character with neutral valence. This might be because, in the context of a hypothetical survival scenario, people might ascribe more trustworthiness and persuasiveness to an agent who appears 'serious', as a congruent smiling expression could have been interpreted as mockery of the situation. However, in the case of the audio-visual experiment, even too "serious" agents were not trusted, perhaps because this was perceived as a lack of empathy towards the human—who is the only one who would have perished on the moon, had the mission failed—hence the preference for a middle ground. As one participant in Experiment 2, $V_n$ condition, remarked, "I didn't like the avatar because he worded his responses in plural second-person (we, us) even though we both know that my life was really the only one at risk".

Although limited, our results expand on previous studies investigating the effect of emotional expressions on social decision-making. Several studies looked at how people reacted to avatars expressing emotions that were incongruent with their behaviour [4, 27, 28], and found extensive evidence that people cooperate more with a congruent avatar (positive emotional expression + cooperative behaviour) than an incongruent one (positive emotional expression + non-cooperative behaviour). These results provide evidence for the EASI model [118]. We contribute to this line of work by presenting people with an avatar that expresses an incongruent emotion in the face and in the voice, rather than in the face and in the strategy. Our results also contribute to the research

on mismatched multi-modal emotional expressions, that so far has mainly been conducted with human sources [24, 25, 43, 90]. Studies that employed artificial agent sources in the past focused on emotion recognition [85, 86]. To the best of our knowledge, ours is the first study linking these two strands of research to investigate the effect of mismatched multi-modal emotional expression on behaviour.

We also used a photorealistic avatar that is more advanced than those previously used [86], thus being more representative of what people are likely to encounter in the near future. What is more, we conducted our experiments "in the wild", collecting data from a large sample size. Our participants took part in the study out of interest, and were not drawn by extrinsic motivation such as monetary compensation. Arguably, this means that their behaviour in the survival task was a more accurate representation of how they would normally behave, than if the experiment had taken place in a lab. The fact that there were no researchers overlooking participants also means that they were less likely to have experienced "white coat" biases. We would probably have had more control over participants' behaviour—in terms of paying attention to the task and not getting distracted by the environment—if we had run a standard laboratory experiment; however, participants' behaviour in a lab might not reflect their behaviour in real life, yet it is in real life that they might need to interact with an artificial agent.

Our experiments differed from previous ones also in terms of how we implemented and analysed the "lunar survival task". Some studies created the artificial agent's ranking a priori [e.g., 5], whereas we created it based on participants' initial ranking, so that each participant saw the avatar make exactly the same amount of changes. This is a more complete manipulation, which has also been used in the past [e.g., 1, 83, 87]. Consequently, it resulted in a more complex design, as demonstrated by the object-specific preferences that emerged (Sections 5.4.2, 6.2.2, and 7.2.2). The more complex design, combined with the complex emotional expression manipulation, might have led to the limited effect of the latter. The choice of utterances used by the agent might also have influenced the results. Some previous studies used self-referencing formulas such as "I ranked [object X] because..." [75, 119], whereas we opted for more impersonal utterances, such as "[Object X] would be useful because..." (see Appendix). Self-referencing techniques such as those used in previous studies have been shown to increase persuasion [14], so lacking this tactic might have reduced the overall persuasiveness of our agent.

### 8.2 Effect of Channel Mismatch

We mismatched the emotional expression in the face and voice of the avatar in order to investigate if the hypothesised "emotional McGurk effect" [43] influenced behaviour, rather than emotion recognition [e.g., 86]. Previous studies found differences in emotion recognition when both the audio and visual channels were present, as opposed to only one. For example, Mower et al. [86] found that emotion classification rate—in terms of arousal, valence, and dominance—was highest for audio-visual and audio only channels, and lowest for video only.

In terms of behavioural reactions, we found more similarities between the audio-visual and the visual only versions of the experiment, suggesting that people might infer the same type of emotional expressivity, and act accordingly, whenever the visual modality is present. However, we also found different behavioural results in the three museum experiments. This suggests that the combined audio-visual modality is not simply made up by the sum of the two individual audio and video channels, but rather it derives from an interaction of these two channels [see 86]. It is possible that a combination of positive valence and low activation gave an impression of seriousness and appropriateness in the audio-visual modality; on the other hand, when one channel was not available, people had to infer the valence and activation information from a more limited source.

From the online study we saw that people indeed extracted valence and arousal information from the video only and audio only channels.

Interestingly, the valence and arousal ratings reveal that the $V_nF_s$ condition was perceived almost identically to the $V_sF_s$ condition (Figures 7, 8); yet, in Experiment 1, we find some evidence that people behaved differently in the $V_nF_s$ condition than the $V_sF_s$ condition. This highlights the importance of conducting behavioural experiments, since often people's ratings do not correlate with actual behaviour [26, 50].

From the online study we also found that, when audio-visual information is available, only information from the facial channel is used to infer the valence ratings; this is consistent with the idea of an "emotional McGurk effect", according to which the audio channel is used to infer activation, and the video channel to infer valence [43, 54]. Mower et al. [86] suggested that this phenomenon can be framed as an information allocation problem: the emotional information to be transmitted is determined by the bandwidth of the available channels. In the case of audio-visual modalities, people have more available information to perceive a certain emotion, similarly to having more contextual information [57]. When there is only one modality, this information is reduced, and the message needs to be perceived from incomplete sources.

### 8.3 Perceptual Questionnaires

We also collected explicit ratings of the virtual character. In the first museum experiment, we found that the agents that were smiling in the face (conditions $V_nF_s$ and $V_sF_s$) were rated as happier, suggesting that people used visual cues, rather than auditory cues, to infer happiness. The $V_nF_s$ agent was also rated as more trustworthy, appealing, and knowledgeable than the other conditions. Factor analysis suggested that these traits can be grouped in two: knowledgeableness, intelligence, trustworthiness belong to one group, and appeal, attractiveness, realism, eeriness, and happiness to another.

Notably, the condition $V_nF_s$ scored higher in the first group of traits. Here, behavioural results partially support perceptual results, as participants accepted suggestions to move the objects from the agent in the $V_nF_s$ condition more. Also, perceiving the agent as intelligent/knowledgeable/trustworthy increased the overall agent's social influence score, while perceiving it as happy decreased it. As previously discussed, happiness might not be the most appropriate emotion to display in a disaster scenario, especially when the other's life is the only one at risk. Van Kleef et al. [118] proposed that the meaning of emotional expressions is mediated by the cooperative or competitive context in which they appear, but this context can likely be extended to positive or negative situations more broadly. Just like laughing at a funeral is inappropriate in some cultures, smiling in a disaster scenario is probably inappropriate too [cf. 56, 57].

In the second experiment (audio-only), in contrast to the audio-visual experiment, there were not many differences in the perceptual evaluations of the two agent conditions ($V_s$ and $V_n$). Marginal effects were found only for attractiveness, with $V_s$ being rated as more attractive than $V_n$. This is in line with previous studies on facial attractiveness, whereby smiling faces are consistently rated as more attractive [65, 96]. To the best of our knowledge, there are no studies that have examined the perceived attractiveness of smiling and neutral voices, but from the current results, it seems that smiling-induced attractiveness extends to smiling voices.

Factor analysis suggested that the perceptual traits could be grouped in 2: attractiveness, appeal, realism, trustworthiness, happiness in one group; and knowledgeableness, intelligence, eeriness in the other. Interestingly, these loadings were different from the audio-visual experiment, as the trustworthiness and eeriness traits switched place. However, none of the emotional expression conditions scored higher in either of the two groups. In the survival task, people accepted more

suggestions to move some objects from the neutral voice than the smiling voice; furthermore, they accepted more suggestions if they perceived the agent to be trustworthy and intelligent.

In the third experiment (video-only), similarly to Experiment 2, there were not many differences in the perceptual evaluation of the agent exhibiting the two emotional expressions ($F_s$ and $F_n$); there were only marginal effects for happiness, with people perceiving $F_s$ as happier than $F_n$. Interestingly then, the previous findings that smiling faces are considered more attractive [65, 96] are not replicated here, as we found no differences in the attractiveness ratings for the $F_s$ and $F_n$ conditions. There are several possible explanations for this result. It is possible that the intensity of the smiling in the voice and face of our agent was perceptually different, so that it was associated with different perceptual traits. Both audio and video channels were obtained from the same person, who could not have produced facial and vocal smiles of different intensities simultaneously; however, as previously mentioned (Section 3), in order to maximise lip-syncing consistency, we dubbed all the audio on videos coming from different recording takes. Thus, it is possible that one take might have been more intense than another, although arguably all these takes would have been more "smiling" than the neutral takes. Another explanation is that the mediation of the disaster scenario context dampened the effect of smiling on facial attractiveness more than on vocal attractiveness. It is also possible that people simply did not find our avatar attractive. In the two experiments where the agent's face was visible, it was rated low in terms of attractiveness, across all experimental conditions (mean for Experiment 1 = 1.79, mean for Experiment 3 = 1.62). On the other hand, in Experiment 2 the average ratings for vocal attractiveness were higher (mean = 2.39). As both the visual appearance and the voice of the character were constant in the three experiments, we can exclude any confounds due to attractiveness in the behavioural results. Anecdotal evidence from participants' comments suggests that people did not like the fact that the avatar was bald, which could explain the low attractiveness score in conditions where this feature was visible. However, future research could focus on the interplay between facial and vocal attractiveness and emotional expressivity, for example, by pre-screening agent's faces and voices for perceived attractiveness.

Factor analysis showed that the traits in Experiment 3 can be grouped in two: knowledgeableness, intelligence, trustworthiness in one group, and appeal, realism, attractiveness, happiness, and eeriness in the other group. Interestingly, these are the same factor loadings as the audio-visual experiment. However, like in the audio-only experiment, none of the emotional expressions scored higher on any of the two groups, although perceiving the agent's face as trustworthy and knowledgeable increased the acceptance of its recommendations, while perceiving it as eerie, or if it was smiling, decreased it.

Overall, the two separate groups of factor loadings seem to be divided in traits related to competence and experience (knowledge, intelligence, trustworthiness), and traits related to appearance (attractiveness, realism, appeal, eeriness). These two groups of traits are usually interconnected, as e.g., attractive people tend to be also perceived as trustworthy, and trusted more [e.g., 15, 109, 120]. Interestingly, however, our participants mapped these traits as perceptually belonging to two different groups, as evinced by the separate factor loadings in the analysis. It is possible that the connection between appearance and competence does not hold when what is judged is an artificial agent, rather than a human; in fact, performance was found to be one of the most effective predictors of trust in Human–Robot Interaction [53]. Biologically speaking, humans tend to attribute more positive traits to individuals who look "good" because looking "good" is a sign of good health and, as a consequence, higher chance of reproductive success [e.g., 48, 91, 98]. However, these appearance-based attributions in an artificial agent might play a minor role compared to competence-based attributions.

### 8.4 Individual Differences

We found little variation in terms of individual differences: an effect of English language fluency in Experiment 1, and an effect of age and gender in Experiment 2. Given the wide variety of participants in our sample, in terms of gender, age, and geographic origin, it is striking that we did not find many more differences at the individual level. Thus, it seems that people's behaviour when interacting with a virtual character is rather generalisable.

We collected data from people coming from all sorts of cultural backgrounds. While smiling as an emotional expression is universal across cultures [29, 37], there is evidence that the perception of smiles, for example, in terms of appropriateness, varies significantly [49, 101, 102, 110]. For example, it was found that people from the USA find it appropriate to smile at higher-status interlocutors, while people from China do not [121]. Thus, it is possible that cultural differences—which would not emerge from the country of origin data that we collected—might have added unexplained variance to the results. Indeed, excluding basic speakers of English from Experiment 1 resulted in perceived avatar happiness being no longer a predictor of social influence. Thus, it is possible that those participants reflected cultural norms according to which a display of happiness is not appropriate in a disaster scenario, or for an avatar in general.

The effect of English language fluency —with the avatar having higher social influence on more fluent speakers— might also be explained by homophily effects: we tend to perceive people who speak our same language as more effective communicators and as more credible [68, 99]. Furthermore, we generally consider people who are more similar to us as more trustworthy [e.g., 46]. Thus, people who perceived the avatar as more similar to themselves, for example, in terms of speaking the same language, might have accepted its recommendations more.

### 8.5 Object Preference

In the first experiment, where people interacted with a virtual agent that had both a face and a voice, we found that people were more easily persuaded to move some objects than others (such as milk more than heater at position #1). In general, participants were more likely to place the milk and transmitter in top positions, and the parachute and raft in bottom position (Figure 18). In the case of milk, people who chose it as their first choice overall made smaller changes in their ranking as compared to, for example, people who had chosen the parachute as their first choice. However, people seemed to have no problem in demoting their other favourite object, the transmitter, when asked by the agent. Thus, while people might have stronger opinions about some specific objects, the same cannot be generalised to all objects.

In the second experiment, where people interacted with the agent only through its voice, we again found that people's favourite objects were the transmitter and milk (Figure 18), and people's least favourite objects were the parachute and raft. In contrast to the audio-visual experiment, however, there were not many differences in terms of which objects people were more persuaded to move, with some marginal exceptions when the object had originally been placed by participants at position #1. Additionally, the overall behaviour towards the agent was not influenced by which object people had placed in the 4 positions—differing from the audio-visual experiment, where people who had placed milk at position #1 were less persuaded by the avatar.

In the third experiment, where people interacted with a virtual character that had only a face, we found people's favourite objects to be the same as the other two experiments (transmitter was still the favourite object, parachute was still the least favourite one, Figure 18). Here we found a few differences on what objects were most likely to be moved, all involving milk, like in Experiment 1. However, the overall behaviour of participants was not influenced by which object they had placed in the four positions.

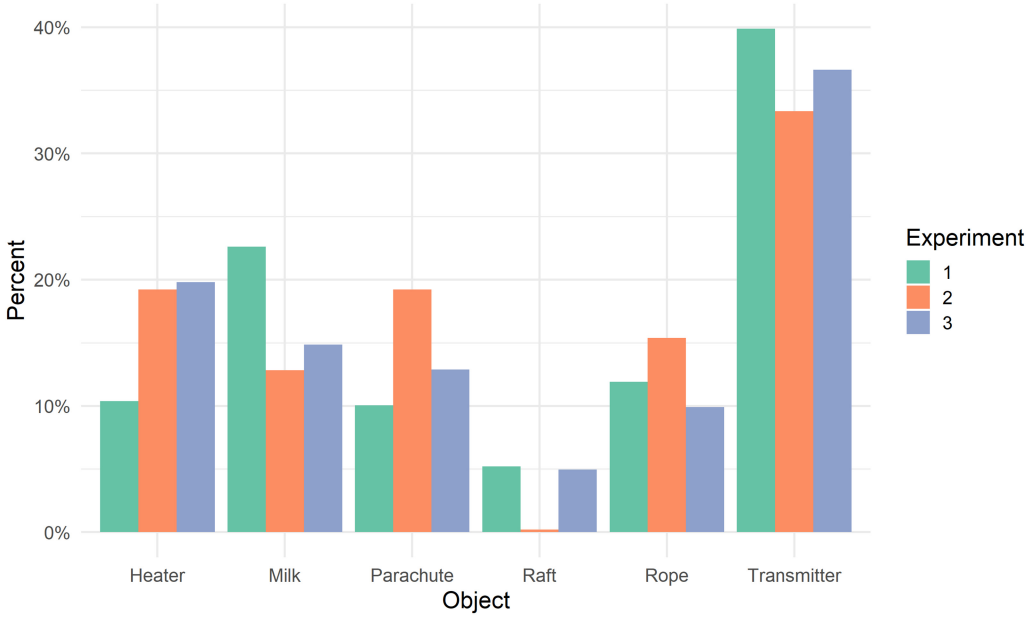


Fig. 18. Percentage of people who originally placed a certain object at position #1 in all three experiments.

Thus, although we had expected that people's object rankings would follow a unimodal distribution, it seems that people had stronger preferences for certain objects (Figure 18), and that they could be less persuaded to change their position. So, while emotional expressivity influenced people's compliance in some cases, in others it seems that the strength of the participants' beliefs, and the content of the agent's argument might have been more influential [cf 53].

We also compared people's willingness to accept the avatar's recommendation with how "correct" their initial rankings were. Interestingly, in Experiment 1 we found that people whose ratings were more similar to the experts' were also more likely to follow the avatar's suggestions. This is the opposite of what one would expect: participants who rank the objects more similarly to the NASA experts probably have some knowledge on the topic, and therefore should be less convinced by the avatar's arbitrary reasoning. However, in our experiment we did not use the full list of 15 objects provided by NASA, but only 6 that were neither on the top nor bottom of the NASA ranking. Thus, since none of these objects have much absolute importance, people might have been more willing to accept an alternative point of view. It might also be that people who are themselves more knowledgeable are more willing to listen to new evidence, in a sort of Dunning–Kruger effect [36]. Of course, we do not have much information on the five NASA experts who created the baseline rankings [52], but it is unlikely that they actually experienced being stranded on the moon with those objects. Thus, the expert ranking might not be much more expert, or meaningful, than those of a population sample who visits science museums.

### 8.6 Limitations and Future Work

There are some limitations that should be mentioned. First of all, our experiment had a rather complex design, partly due to the nature of the survival task, partly to our implementation of it, for example, in terms of providing an agent ranking based on participants' initial ranking. This allowed us to maintain control over how much the avatar's ranking differed from the participants',

but prevented us from calculating simpler measurements, such as initial and final divergence from the avatar [74].

Second, our assumption that people would be equally likely to rate any of the six objects as most important was proven wrong. This forced us to conduct many follow-up analyses to find any behavioural patterns that might have been masked by people's subjective object preferences. The substantial power afforded by our large sample size allowed us to uncover such patterns, but the nature of the results remains exploratory.

Thirdly, as shown in the online study, the vocal and facial smiling components of the audio-visual stimuli were perceived differently in terms of arousal and valence. Specifically, while smiling in the face increased ratings of both arousal and valence, smiling in the voice only increased ratings of arousal. We have already discussed how this supports previous hypotheses on the "emotional McGurk effect". However, this also means that there might have been an asymmetry in the perception—and behavioural reaction—of these two components during the "lunar survival task".

Also, as the "lunar survival task" is language-based, and as our participant sample comprised people from a wide geographic background, it is possible that not everyone understood the avatar's speech in the same manner. While we partially accounted for this using participants' self-reported English language fluency as a predictor in our analyses, we cannot rule out the possibility that some participants might have played the game without understanding what it exactly entailed.

Another consideration is that we used only one avatar. Although our emotional expressions were validated both by means of categorical labels (happiness ratings at the end of the experiments) and dimensional ratings of valence and arousal (online evaluation study), future studies should look at how the avatar's visual appearance interacts with its emotional expressivity. As we saw, people did not find our avatar attractive. It would be interesting to compare our avatar with a more attractive one. It would also be interesting to compare an equally attractive, female, avatar, and an avatar differing in terms of realism, to examine potential effects related to the Uncanny Valley.

We also want to continue working on the idea of a "smiling McGurk effect". While our crowd-sourced study showed that people perceived the smiling in the face and voice of our avatar, it is possible that this mismatch will manifest itself more on perception than on decision-making. For example, it would be interesting to see if an avatar showing this emotional mismatch interferes with language or emotion processing [17, 80].

## 9 CONCLUSION

In summary, we conducted four experiments: an online emotion processing experiment, where participants rated an artificial agent that displayed different emotional expressions in its face and voice in terms of valence and arousal; and three experiments in a naturalistic environment, where participants played a "lunar survival task" with the agent. The complex design prevented us from finding any clear-cut results on the influence of emotional expression on decision-making. Substantial follow-up analyses showed that changing the emotional expressivity of an artificial agent in both the voice and the face, or in only one channel, results in different behaviours and perceptions on the part of the human. This lends support to previous theories of audio-visual integration in emotion processing [43, 54, 86]. Specifically, while previous studies dealt with the "Emotional McGurk Effect" of emotion recognition, here we concentrated on the influence that this effect could have on human behaviour. Notably, our findings have implications for Human–Machine Interaction and machine design, for example, in cases where prioritisation of which expressive channel to build is required. With our results, we also want to highlight the importance of the context in which the Human–Machine Interaction takes place. Situational context (our lunar crash scenario) but also linguistic context (the content of the agent's utterances) likely interacted with the agent's emotional expressivity and its social influence. Thus, we would encourage other

researchers and machine designers to think about the emotional valence and arousal for each expressive modality of an artificial agent, in conjunction with the interaction context.

## APPENDIX

### A LIST OF UTTERANCES

Positive descriptions:

(1) I think the receiver transmitter would be important since it's our only means of communication.
(2) I think the rope would be useful for climbing rocks and tying us together so as to not float away.
(3) I think the parachute fabric could protect us from the sun rays.
(4) The gas inside the raft could be used for propulsion, for moving faster.
(5) I think the milk would be a good source of calories and energy.
(6) I think we would need the heater so as not to freeze when we rest.

Neutral descriptions:

(1) The receiver transmitter might be useful, but only if someone happens to be near us.
(2) The rope could be handy, but carrying it would slow us down.
(3) I think we could move faster with the parachutes, but we also risk being blown away.
(4) We could use the raft for protection, but it won't help us advance.
(5) I think the milk could feed us, but it would be bulky to carry.
(6) I think the heater could be useful, but only if we went into the dark side of the moon.

Negative descriptions:

(1) I don't think the receiver transmitter would be useful, since it only works on short distances.
(2) I don't think the rope would help us, it would only impede our movements.
(3) I don't think we would need the parachutes, since we have already landed.
(4) I think the raft would be useless, since there are no water courses on the moon.
(5) I don't think we would need the milk, it would be too much effort to carry it, for too little energy.
(6) The heater would be useless in the direct sun rays exposure.

## STATEMENT OF PREVIOUS RESEARCH

This article presents the full set of analyses of the experiment introduced in Torre et al. [114]. However, here we cover significantly more material, as we report two additional experiments that had not been run yet at the time of our previous article. Also, our previous article only reported preliminary results in terms of number of participants and also in terms of depth of statistical analyses. The current article has not been previously submitted anywhere else.

We have recently submitted another article to ICMI 2019, presenting a new experiment in which we used the same "survival task" and artificial agent. However, this is an entirely different experiment, which was run several months after the experiments presented here, and which deals with the agent's appearance as an experimental condition. Thus, the current article and the article currently under review at ICMI are completely unrelated.

## ACKNOWLEDGMENTS

We are grateful to the Science Gallery staff and visitors.

## REFERENCES


[1] Sigurdur O. Adalgeirsson and Cynthia Breazeal. 2010. MeBot: A robotic platform for socially embodied presence. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE Press, 15–22.

[2] Pierre Y. Andrews. 2012. System personality and persuasion in human-computer dialogue. *ACM Transactions on Interactive Intelligent Systems* 2, 2 (2012), 12.

[3] Pierre Y. Andrews and Suresh Manandhar. 2009. Measure of belief change as an evaluation of persuasion. In *Proceedings of the Persuasive Technology and Digital Behaviour Intervention Symposium*.

[4] Dimitrios Antos, Celso M. De Melo, Jonathan Gratch, and Barbara J. Grosz. 2011. The influence of emotion expression on perceptions of trustworthiness in negotiation. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*.

[5] Ron Artstein, David Traum, Jill Boberg, Alesia Gainer, Jonathan Gratch, Emmanuel Johnson, Anton Leuski, and Mikio Nakano. 2017. Listen to my body: Does making friends help influence people? In *Proceedings of the 13th International Florida Artificial Intelligence Research Society Conference*.

[6] Véronique Aubergé and Marie Cathiard. 2003. Can we hear the prosody of smile? *Speech Communication* 40, 1 (2003), 87–97.

[7] Annette Baier. 1986. Trust and antitrust. *Ethics* 96, 2 (1986), 231–260. DOI: https://doi.org/10.1086/292745

[8] Moshe Bar, Maital Neta, and Heather Linz. 2006. Very first impressions. *Emotion* 6, 2 (2006), 269–278. DOI: https://doi.org/10.1037/1528-3542.6.2.269

[9] Russell Beale and Chris Creed. 2009. Affective interaction: How emotional agents affect users. *International Journal of Human-Computer Studies* 67, 9 (2009), 755–776.

[10] Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, reciprocity, and social history. *Games and Economic Behavior* 10, 1 (1995), 122–142.

[11] Gerd Bohner, Amanda Dykema-Engblade, R. Scott Tindale, and Helen Meisenhelder. 2008. Framing of majority and minority source information in persuasion: When and how "consensus implies correctness". *Social Psychology* 39, 2 (2008), 108–116.

[12] R. Thomas Boone and Ross Buck. 2003. Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation. *Journal of Nonverbal Behavior* 27, 3 (2003), 163–182.

[13] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49–59.

[14] Robert E. Burnkrant and H. Rao Unnava. 1989. Self-referencing: A strategy for increasing processing of message content. *Personality and Social Psychology Bulletin* 15, 4 (1989), 628–638.

[15] Danilo Bzdok, Robert Langner, Svenja Caspers, Florian Kurth, Ute Habel, Karl Zilles, Angela R. Laird, and Simon B. Eickhoff. 2011. ALE meta-analysis on facial judgments of trustworthiness and attractiveness. *Brain Structure and Function* 215, 3–4 (2011), 209–223.

[16] Angelo Cafaro, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2016. First impressions in human–agent virtual encounters. *ACM Transactions on Computer-Human Interaction* 23, 4 (2016), 1–40.

[17] Salvatore Campanella and Pascal Belin. 2007. Integrating face and voice in person perception. *Trends in Cognitive Sciences* 11, 12 (2007), 535–543.

[18] Vijay Chidambaram, Yueh-Hsuan Chiang, and Bilge Mutlu. 2012. Designing persuasive robots: How robots might persuade people using vocal and nonverbal cues. In *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 293–300. DOI: https://doi.org/10.1145/2157689.2157798

[19] R. H. B. Christensen. 2015. ordinal—Regression Models for Ordinal Data. R package version 2015.6-28. Retrieved from http://www.cran.r-project.org/package=ordinal/.

[20] Martin D. Coleman. 2011. Emotion and the self-serving bias. *Current Psychology* 30, 4 (2011), 345–354.

[21] John Cook and Toby Wall. 1980. New work attitude measures of trust, organizational commitment and personal need non-fulfilment. *Journal of Occupational Psychology* 53, 1 (1980), 39–52.

[22] Alan S. Cowen, Petri Laukka, Hillary Anger Elfenbein, Runjing Liu, and Dacher Keltner. 2019. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature Human Behaviour* 3, 4 (2019), 369–382.

[23] Chris Creed and Russell Beale. 2008. Psychological responses to simulated displays of mismatched emotional expressions. *Interacting with Computers* 20, 2 (2008), 225–239.

[24] Beatrice de Gelder, Koen B. E. Böcker, Jyrki Tuomainen, Menno Hensen, and Jean Vroomen. 1999. The combined perception of emotion from voice and face: Early interaction revealed by human electric brain responses. *Neuroscience Letters* 260, 2 (1999), 133–136.

[25] Beatrice De Gelder and Jean Vroomen. 2000. The perception of emotions by ear and by eye. *Cognition & Emotion* 14, 3 (2000), 289–311.

[26] Jan De Houwer. 2006. What are implicit measures and why are we using them. In *Proceedings of the Handbook of Implicit Cognition and Addiction*. R. W. Wiers and A. W. Stacy (Eds.) Sage Publications, 11–28.

[27] Celso M. de Melo, Peter Carnevale, and Jonathan Gratch. 2013. People's biased decisions to trust and cooperate with agents that express emotions. In *Proceedings from International Conference on Autonomous Agents and Multiagent Systems*.
[28] Celso M. de Melo, Jonathan Gratch, and Peter J. Carnevale. 2015. Humans versus computers: Impact of emotion expressions on people's decision making. *IEEE Transactions on Affective Computing* 6, 2 (2015), 127–136.
[29] George V. N. Dearborn. 1900. The nature of the smile and laugh. *Science* 11, 283 (1900), 851–856.
[30] Bella M. DePaulo, Amy L. Blank, Gregory W. Swaim, and Joan G. Hairfield. 1992. Expressiveness and expressive control. 18, 3 (1992), 276–285. DOI: https://doi.org/10.1177/0146167292183003
[31] David DeSteno, Cynthia Breazeal, Robert H. Frank, David Pizarro, Jolie Baumann, Leah Dickens, and Jin Joo Lee. 2012. Detecting the trustworthiness of novel partners in economic exchange. *Psychological Science* 23, 12 (2012), 1549–1556.
[32] Matthieu Destephe, Martim Brandao, Tatsuhiro Kishi, Massimiliano Zecca, Kenji Hashimoto, and Atsuo Takanishi. 2015. Walking in the uncanny valley: Importance of the attractiveness on the acceptance of a robot as a working partner. *Frontiers in Psychology* 6 (2015), 204.
[33] Karen Dion, Ellen Berscheid, and Elaine Walster. 1972. What is beautiful is good. *Journal of Personality and Social Psychology* 24, 3 (1972), 285.
[34] Amy Drahota, Alan Costall, and Vasudevi Reddy. 2008. The vocal communication of different kinds of smile. *Speech Communication* 50, 4 (2008), 278–287. DOI: https://doi.org/10.1016/j.specom.2007.10.001
[35] Jennifer R. Dunn and Maurice E. Schweitzer. 2005. Feeling and believing: The influence of emotion on trust. *Journal of Personality and Social Psychology* 88, 5 (2005), 736.
[36] David Dunning. 2011. The Dunning–Kruger effect: On being ignorant of one's own ignorance. In *Proceedings of the Advances in Experimental Social Psychology*. Vol. 44. Elsevier, 247–296.
[37] Irenäus Eibl-Eibesfeldt. 1972. Similarities and differences between cultures in expressive movements. In *Proceedings of the Non-verbal Communication*, R. A. Hinde (Ed.). Cambridge University Press.
[38] Hedwig Eisenbarth and Georg W. Alpers. 2011. Happy mouth and sad eyes: Scanning emotional facial expressions. *Emotion* 11, 4 (2011), 860.
[39] Paul Ekman and Wallace V. Friesen. 1982. Felt, false, and miserable smiles. *Journal of Nonverbal Behavior* 6, 4 (1982), 238–252.
[40] Kevin El Haddad, Ilaria Torre, Emer Gilmartin, Hüseyin Çakmak, Stéphane Dupont, Thierry Dutoit, and Nick Campbell. 2017. Introducing AmuS: The amused speech database. In *Proceedings of the International Conference on Statistical Language and Speech Processing*. Springer, 229–240.
[41] Hillary Anger Elfenbein and Nalini Ambady. 2002. On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin* 128, 2 (2002), 203.
[42] Aaron C. Elkins and Douglas C. Derrick. 2013. The sound of trust: Voice as a measurement of trust during interactions with embodied conversational agents. *Group Decision and Negotiation* 22, 5 (2013), 897–913.
[43] Sascha Fagel. 2006. Emotional McGurk effect. In *Proceedings of the International Conference on Speech Prosody*, Vol. 1. Dresden, Germany.
[44] Sascha Fagel. 2009. Effects of smiled speech on lips, larynx and acoustics. In *Proceedings of the Auditory Visual Speech Processing Conference*. Norwich, UK, 18–21.
[45] Agneta H. Fischer, Lisanne S. Pauw, and Antony S. R. Manstead. 2019. Emotion recognition as a social act: The role of the expresser-observer relationship in recognizing emotions. In *Proceedings of the Social Nature of Emotion Expression*, Ursula Hess and Shlomo Hareli (Eds.). Springer, 7–24.
[46] Margaret Foddy, Michael J. Platow, and Toshio Yamagishi. 2009. Group-based trust in strangers: The role of stereotypes and expectations. *Psychological Science* 20, 4 (2009), 419–422.
[47] Robert H. Frank. 1988. *Passions Within Reason: The Strategic Role of the Emotions*. WW Norton & Co.
[48] Steven W. Gangestad and Glenn J. Scheyd. 2005. The evolution of human physical attractiveness. *Annual Review of Anthropology* 34, 1 (2005), 523–548.
[49] Jeffrey M. Girard and Daniel McDuff. 2017. Historical heterogeneity predicts smiling: Evidence from large-scale observational analyses. In *Proceedings of the 12th International Conference on Automatic Face & Gesture Recognition*. IEEE, 719–726.
[50] Anthony G. Greenwald. 1990. What cognitive representations underlie social attitudes? *Bulletin of the Psychonomic Society* 28, 3 (1990), 254–260.
[51] Sarah D. Gunnery, Judith A. Hall, and Mollie A. Ruben. 2013. The deliberate Duchenne smile: Individual differences in expressive control. *Journal of Nonverbal Behavior* 37, 1 (2013), 29–41.
[52] Jay Hall and Wilfred Harvey Watson. 1970. The effects of a normative intervention on group decision-making performance. *Human Relations* 23, 4 (1970), 299–317.

[53] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53, 5 (2011), 517–527.

[54] Alan Hanjalic. 2006. Extracting moods from pictures and sounds: Towards truly personalized TV. *IEEE Signal Processing Magazine* 23, 2 (2006), 90–100.

[55] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1–14.

[56] Ursula Hess, Jonas Dietrich, Konstantinos Kafetsios, Shimon Elkabetz, and Shlomo Hareli. 2020. The bidirectional influence of emotion expressions and context: Emotion expressions, situational information and real-world knowledge combine to inform observers' judgments of both the emotion expressions and the situation. *Cognition and Emotion* 34, 3 (2020), 539–552.

[57] Ursula Hess and Shlomo Hareli. 2017. The social signal value of emotions: The role of contextual factors in social inferences drawn from emotion displays. In *Proceedings of the Oxford Series in Social Cognition and Social Neuroscience. The Science of Facial Expression.*, J.-M. Fernández-Dols & J. A. Russell (Ed.). Oxford University Press, 375–393.

[58] Dacher Keltner and Jonathan Haidt. 1999. Social functions of emotions at four levels of analysis. *Cognition & Emotion* 13, 5 (1999), 505–521.

[59] Peter Khooshabeh, Celso M. De Melo, Brooks Volkman, Jonathan Gratch, Jim Blascovich, and Peter Carnevale. 2013. Negotiation strategies with incongruent facial expressions of emotion cause cardiovascular threat. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 35.

[60] Peter Khooshabeh, Cade McCall, Sudeep Gandhe, Jonathan Gratch, and James Blascovich. 2011. Does it matter if a computer jokes? In *Proceedings of CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, 77–86.

[61] Andrea Kleinsmith and Nadia Bianchi-Berthouze. 2013. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing* 4, 1 (2013), 15–33.

[62] Eva Krumhuber, Antony S. R. Manstead, Darren Cosker, Dave Marshall, Paul L. Rosin, and Arvid Kappas. 2007. Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion* 7, 4 (2007), 730–735.

[63] Marianne LaFrance and Marvin A. Hecht. 1995. Why smiles generate leniency. *Personality and Social Psychology Bulletin* 21, 3 (1995), 207–214.

[64] Peter J. Lang, Mark K. Greenwald, Margaret M. Bradley, and Alfons O. Hamm. 1993. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology* 30, 3 (1993), 261–273.

[65] Sing Lau. 1982. The effect of smiling on person perception. *The Journal of Social Psychology* 117, 1 (1982), 63–67. DOI: https://doi.org/10.1080/00224545.1982.9713408

[66] Iolanda Leite, André Pereira, Carlos Martinho, and Ana Paiva. 2008. Are emotional robots more fun to play with? In *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium On*. IEEE, 77–82.

[67] Russell Lenth. 2020. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. Retrieved from https://CRAN.R-project.org/package=emmeans. R package version 1.4.7.

[68] Shiri Lev-Ari and Boaz Keysar. 2010-11. Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology* 46, 6 (2010–11), 1093–1096. DOI: https://doi.org/10.1016/j.jesp.2010.05.025

[69] Robert W. Levenson. 1994. Human emotions: A functional view. In *Proceedings of the Nature of Emotion*, Davidson R. Ekman, Paul (Ed.). Oxford University Press, 123–126.

[70] Jamy Li, Wendy Ju, and Clifford I. Nass. 2015. Observer perception of dominance and mirroring behavior in human-robot relationships. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 133–140.

[71] Benny Liebold and Peter Ohler. 2013. Multimodal emotion expressions of virtual agents, mimic and vocal emotion expressions and their effects on emotion recognition. In *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 405–410.

[72] Robert B. Lount Jr. 2010. The impact of positive mood on trust in interpersonal and intergroup interactions. *Journal of Personality and Social Psychology* 98, 3 (2010), 420.

[73] Gale Lucas, Giota Stratou, Shari Lieblich, and Jonathan Gratch. 2016. Trust me: Multimodal signals of trustworthiness. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 5–12.

[74] Gale M. Lucas, Jill Boberg, David Traum, Ron Artstein, Jonathan Gratch, Alesia Gainer, Emmanuel Johnson, Anton Leuski, and Mikio Nakano. 2018. Getting to know each other: The role of social dialogue in recovery from errors in social robots. In *Proceedings of the 13th Annual ACM/IEEE International Conference on Human-Robot Interaction*. 344–351.

[75] Gale M. Lucas, Janina Lehr, Nicole Krämer, and Jonathan Gratch. 2019. The effectiveness of social influence tactics when used by a virtual agent. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 22–29.
[76] Maya B. Mathur and David B. Reichling. 2016. Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition* 146 (2016), 22–32.
[77] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *The Academy of Management Review* 20, 3 (1995), 709–734. DOI: https://doi.org/10.2307/258792
[78] Phil McAleer, Alexander Todorov, and Pascal Belin. 2014. How do you say hello? Personality impressions from brief novel voices. *PLoS ONE* 9, 3 (2014), e90779.
[79] Robert R. McCrae. 2009. The five-factor model of personality. In *Proceedings of the Handbook of Personality Psychology*, P. Corr and G. Mathhews (Eds.). 148–161.
[80] Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264, 5588 (1976), 746.
[81] Marc Mehu and Robin I. M. Dunbar. 2008. Relationship between smiling and laughter in humans (Homo sapiens): Testing the power asymmetry hypothesis. *Folia Primatologica* 79, 5 (2008), 269–280.
[82] Laura Mieth, Raoul Bell, and Axel Buchner. 2016. Facial likability and smiling enhance cooperation, but have no direct effect on moralistic punishment. *Experimental Psychology* 63, 5 (2016), 263–277.
[83] Christine Moon, Robin Panneton Cooper, and William P. Fifer. 1993. Two-day-olds prefer their native language. *Infant Behavior and Development* 16, 4 (1993), 495–500.
[84] Youngme Moon. 1998. The effects of distance in local versus remote human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 103–108.
[85] Emily Mower, Sungbok Lee, Maja J. Mataric, and Shrikanth Narayanan. 2008. Human perception of synthetic character emotions in the presence of conflicting and congruent vocal and facial expressions. In *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2201–2204.
[86] Emily Mower, Maja J. Mataric, and Shrikanth Narayanan. 2009. Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information. *IEEE Transactions on Multimedia* 11, 5 (2009), 843–855.
[87] Clifford I. Nass, B. J. Fogg, and Youngme Moon. 1996. Can computers be teammates? *International Journal of Human-Computer Studies* 45, 6 (1996), 669–678.
[88] Thierry Nazzi, Josiane Bertoncini, and Jacques Mehler. 1998. Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance* 24, 3 (1998), 756.
[89] Ian S. Penton-Voak, Nicholas Pound, Anthony C. Little, and David I. Perrett. 2006. Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" In Social Perception. *Social Cognition* 24, 5 (2006), 607–640. DOI: https://doi.org/10.1521/soco.2006.24.5.607
[90] Gilles Pourtois, Beatrice de Gelder, Anne Bol, and Marc Crommelinck. 2005. Perception of facial expressions and voices and of their combination in the human brain. *Cortex* 41, 1 (2005), 49–59.
[91] Pavol Prokop and Peter Fedor. 2011. Physical attractiveness influences reproductive success of modern men. *Journal of Ethology* 29, 3 (2011), 453.
[92] R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.
[93] Robin Read and Tony Belpaeme. 2012. How to use non-linguistic utterances to convey emotion in child-robot interaction. In *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 219–220.
[94] Lawrence Ian Reed, Rachel Stratton, and Jessica D. Rambeas. 2018. Face value and cheap talk: How smiles can increase or decrease the credibility of our words. *Evolutionary Psychology* 16, 4 (2018), 1474704918814400.
[95] Christina Regenbogen, Daniel A. Schneider, Andreas Finkelmeyer, Nils Kohn, Birgit Derntl, Thilo Kellermann, Raquel E. Gur, Frank Schneider, and Ute Habel. 2012. The differential contribution of facial expressions, prosody, and speech content to empathy. *Cognition & Emotion* 26, 6 (2012), 995–1014.
[96] Harry T. Reis, Ilona McDougal Wilson, Carla Monestere, Stuart Bernstein, Kelly Clark, Edward Seidl, Michelle Franco, Ezia Gioioso, Lori Freeman, and Kimberly Radoane. 1990. What is smiling is beautiful and good. *European Journal of Social Psychology* 20, 3 (1990), 259–267.
[97] William Revelle. 2018. *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. Retrieved from https://CRAN.R-project.org/package=psych. R package version 1.8.12.
[98] Gillian Rhodes. 2006. The evolutionary psychology of facial beauty. *Annual Review of Psychology* 57, 1 (2006), 199–226.
[99] Everett M. Rogers and Dilip K. Bhowmik. 1970. Homophily-heterophily: Relational concepts for communication research. *Public Opinion Quarterly* 34, 4 (1970), 523–538.

[100] Magdalena Rychlowska, Rachael E. Jack, Oliver G. B. Garrod, Philippe G. Schyns, Jared D. Martin, and Paula M. Niedenthal. 2017. Functional smiles: Tools for love, sympathy, and war. *Psychological Science* 28, 9 (2017), 1259–1270.
[101] Magdalena Rychlowska, Antony S. R. Manstead, and Job van der Schalk. 2019. The many faces of smiles. In *Proceedings of the Social Nature of Emotion Expression*, Ursula Hess and Shlomo Hareli (Eds.). Springer, 227–245.
[102] Magdalena Rychlowska, Yuri Miyamoto, David Matsumoto, Ursula Hess, Eva Gilboa-Schechtman, Shanmukh Kamble, Hamdi Muluk, Takahiko Masuda, and Paula Marie Niedenthal. 2015. Heterogeneity of long-history migration explains cultural differences in reports of emotional expressivity and the functions of smiles. *Proceedings of the National Academy of Sciences* 112, 19 (2015), E2429–E2436.
[103] Klaus R. Scherer. 1987. Toward a dynamic theory of emotion: The component process model of affective states. *Geneva Studies in Emotion and Communication* 1 (1987), 1–98.
[104] Klaus R. Scherer, Rainer Banse, and Harald G. Wallbott. 2001. Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-cultural Psychology* 32, 1 (2001), 76–92.
[105] Matthias Scheutz, Paul W. Schermerhorn, and James Kramer. 2006. The utility of affect expression in natural language interactions in joint human-robot tasks. In *Proceedings of the 1st Annual ACM/IEEE International Conference on Human-Robot Interaction*. 226–233.
[106] Joanna Schug, David Matsumoto, Yutaka Horita, Toshio Yamagishi, and Kemberlee Bonnet. 2010. Emotional expressivity as a signal of cooperation. *Evolution and Human Behavior* 31, 2 (2010), 87–94.
[107] Ameneh Shamekhi, Mary Czerwinski, Gloria Mark, Margeigh Novotny, and Gregory A. Bennett. 2016. An exploratory study toward the preferred conversational style for compatible virtual agents. In *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, 40–50.
[108] Herbert A. Simon. 1967. Motivational and emotional controls of cognition. *Psychological Review* 74, 1 (1967), 29–39.
[109] Melissa K. Surawski and Elizabeth P. Ossoff. 2006. The effects of physical and vocal attractiveness on impression formation of politicians. *Current Psychology* 25, 1 (2006), 15–27.
[110] Piotr Szarota. 2010. The mystery of the European smile: A comparison based on individual photographs provided by internet users. *Journal of Nonverbal Behavior* 34, 4 (2010), 249–256.
[111] Sachiko Takagi, Saori Hiramatsu, Ken-ichi Tabei, and Akihiro Tanaka. 2015. Multisensory perception of the six basic emotions is modulated by attentional instruction and unattended modality. *Frontiers in Integrative Neuroscience* 9 (2015), 1. DOI: https://doi.org/10.3389/fnint.2015.00001
[112] Leila Takayama, Victoria Groom, and Clifford I. Nass. 2009. I'm sorry, Dave: I'm afraid I won't do that: Social aspects of human-agent conflict. In *Proceedings of the 2009 SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2099–2108.
[113] Ilaria Torre. 2014. Production and perception of smiling voice. In *Proceedings of the 1st Postgraduate and Academic Researchers in Linguistics at York Conference*. York, U. K.
[114] Ilaria Torre, Emma Carrigan, Killian McCabe, Rachel McDonnell, and Naomi Harte. 2018. Survival at the museum: A cooperation experiment with emotionally expressive virtual characters. In *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 423–427.
[115] 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. DOI: https://doi.org/10.1016/j.chb.2019.106215
[116] 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 APAScience'18: Technology, Mind, and Society*. DOI: https://doi.org/10.1145/3183654.3183691
[117] Gerben A. Van Kleef, Carsten K. W. De Dreu, and Antony S. R. Manstead. 2004. The interpersonal effects of anger and happiness in negotiations. *Journal of Personality and Social Psychology* 86, 1 (2004), 57.
[118] Gerben A. Van Kleef, Carsten K. W. De Dreu, and Antony S. R. Manstead. 2010. An interpersonal approach to emotion in social decision making: The emotions as social information model. *Advances in Experimental Social Psychology* 42 (2010), 45–96.
[119] Yuqiong Wang, Gale Lucas, Peter Khooshabeh, Celso M. De Melo, and Jonathan Gratch. 2015. Effects of emotional expressions on persuasion. *Social Influence* 10, 4 (2015), 236–249.
[120] Rick K. Wilson and Catherine C. Eckel. 2006. Judging a book by its cover: Beauty and expectations in the trust game. *Political Research Quarterly* 59, 2 (2006), 189–202.
[121] Richard L. Wiseman and Xiaohui Pan. 2004. Smiling in the People's Republic of China and the United States: Status and situational influences on the social appropriateness of smiling. *Intercultural Communication Studies* 13 (2004), 1–18.
[122] Katja Zibrek and Rachel McDonnell. 2014. Does render style affect perception of personality in virtual humans? In *Proceedings of the ACM Symposium on Applied Perception*. 111–115.

Received July 2019; revised April 2021; accepted June 2021