# When your Robot Avatar Misbehaves you are Likely to Apologize: an exploration of guilt during robot embodiment

Laura Aymerich-Franch[a, *], Sameer Kishore[a], and Mel Slater[a, b]

[a]EVENT Lab, Department of Personality, Evaluation and Psychological Treatment, Faculty of Psychology, University of Barcelona, 08035 Barcelona, Spain.

[b] Institut de Neurociències, University of Barcelona, 08035 Barcelona, Spain.

[*]Correspondence to: Laura Aymerich-Franch, EventLAB, Universitat de Barcelona, Facultat de Psicologia, Departament de Personalitat, Avaluació i Tractaments Psicològics, Campus de Mundet - Edifici Teatre, Passeig de la Vall d'Hebron 171, 08035 Barcelona, Spain, laura.aymerich@gmail.com, +34 93 403 9618

## Acknowledgments

Laura Aymerich-Franch was supported by the Marie Curie IOF Fellowship project HumRobCooperation, funded by the European Commission, under grant agreement No PIOF-CT-622764. Sameer Kishore was supported by the project PSI2014-56301-R Ser Einstein: La Influencia de Internalizar un Cuerpo Virtual en la Inteligencia, Ministerio de Economía, Industria y Competitividad of Spain.

## Abstract

Would people feel guilty if their robot avatar acted autonomously to harm someone? We examined the experience of guilt during robot avatar embodiment, a form of embodiment where the participants experience the body of a humanoid robot as if it were their own. In particular, we analyzed what happens when a robot avatar spontaneously verbally abuses someone during a conversation using the participant’s voice, without this being the intention of the participant. In a 2×2 between-subjects experimental design, participants embodied a humanoid robot that added either offensive or neutral words during a conversation with a confederate, and had control over the robot’s movements or not (synch. vs. asynch.). We found that guilt and shame were positively associated with offensive words and that apologizing and verbal repair were positively related to guilt. Also, body ownership was moderately associated to apologizing and verbal repair.  The results suggest that people may feel guilty for the actions of their robot avatars even if they are not the real agents of these actions. The work highlights the importance of examining the moral and legal aspects related to robot embodiment technologies.

*Keywords:* robot embodiment, body ownership, avatar, guilt, responsibility, moral emotions, humanoid robots.

# When your Robot Avatar Misbehaves you are Likely to Apologize: an exploration of guilt during robot embodiment

## 1. Introduction

It is likely that robots will become ubiquitous during the next decade. An important use of robotics is for telecommunications, in the form of robot embodiment [1]–[5]. In these systems, people embody a robot in a remote destination to have a physical presence and interact with people there. Being ‘embodied’ means that the participant or ‘visitor’ to the remote place is in a virtual reality system that receives sensory data (vision, sound, touch) from the sensors of the remote robot, so that the participant sees, hears and feels from the cameras (eyes), microphones (ears) and touch sensors on the remote robot. Also, the movements of the participant are tracked in real time, and the data transmitted to the remote robot. Hence movements of the person are reflected in movements of the remote robot. As the participant turns his or her head, so the robot head will turn accordingly, pick up new visual and auditory data, which is then perceived by the participant in almost real time (depending on system latency). During this process, participants experience the robot body as if it were their  own body [6]–[14]. Also, they experience the sense of presence in the location of the robot avatar [15]. It is as if the perceptual consciousness of the participant had been transferred into the body of the remote robot.

One question that arises is what happens when things go wrong? The philosopher Thomas Metzinger [16] has discussed the following scenario: A person, referred to as the visitor, is embodied in a remote robot interacting with people at its location. The robot is controlled by a Brain-Computer Interface (BCI) [12], [17]. A person who in the past has done the visitor great harm walks into the physical scenario where the robot is located. The visitor has a flash of temper, which the BCI interprets as a murderous impulse and kills the person who walked in. Where does the blame lie? Where is the legal and moral responsibility? A momentary thought may not become a legal intention unless the person themselves physically carries out the action. But here the person is embodied as the robot, and to all intents and purposes this may be regarded as his or her body. On the other hand, it is an algorithmic and software fault for the BCI to carry out the murderous action simply because of the spontaneous thought. Apart from the personal responsibility there is also the question of jurisdiction. Is this a crime in the country where the participant is physically located, or where the robot is located? All of these issues will become vital to prepare for and address as this type of technology comes into widespread use [16].

In the present work, we investigated a related issue but one less dramatic than (accidental) murder. In our scenario, participants are embodied in a remote robot, arranged so that they would or would not be likely to have a concomitant illusion of body ownership, and thereby carry out a conversation with a bystander (in fact a confederate). During the conversation, additional words are inserted by the robot as if they were spoken by the participant (and in fact they are in the participant's own voice). These words may be neutral words, or they may be words that insult the remote partner. We carried out an exploratory study to understand how the participant would respond to this situation – where would they put the responsibility, how personal would it be.

More specifically, the present study examines whether participants feel guilty and try to compensate for the insulting actions of their robot avatar when their robot avatar acts antisocially (i.e. verbally offends another person) and these acts are out of their control.

Guilt is an unpleasant although important moral emotion as it allows humans to know when they have harmed someone. In general, after doing something harmful to others, people feel guilty and experience the need to repair damage, for instance, in the form of an apology [18]. It is important to consider though, that not all moral emotions lead to prosocial acts. For instance, [19] found that guilt increases cooperation in dilemma games but shame does not. Only guilt, but not other moral emotions, promote constructive and proactive actions, and leads to reparative actions such as confessions and apologies, as well as to acts intended to undo the consequences of the behavior that caused that emotion [20]–[22]. Guilt arises from the concern for others and is related to actions that are experienced as causing harm [23], [24]. On the other hand, shame arises from the concern with others' evaluations of the self and the failure to meet important personal standards [24], [25]. Shame is related to a global negative evaluation of the self.

Our principal hypothesis was that participants would experience guilt for their robot avatar actions, when they had body ownership over the robot body, even when these were not their fault. While we expected that offensive words pronounced by the robot avatar would lead to an increase both in guilt and shame, guilt is crucial to demonstrate that the participant indeed felt responsible for the actions of the robot avatar when the robot insulted the participant and not only concerned as an external observer could be. Guilt is related to a condemnation of a specific behavior, and occurs when one feels responsible for another's person negative affective state or for harming them. Guilt motivates a heightened sense of personal responsibility [26]. In the frame of the study, we understand sense of responsibility as the self-attributed consideration that one is accountable or to blame for something, which is therefore directly connected to guilt. In our experimental design, it would be possible to experience other moral emotions such as shame for reasons other than harming the participant (e.g. resulting from the fact that the robot used their voice to pronounce non-sense words in between the conversation which made them appear silly in front of the other participant). However, only if participants felt that the other participant was being harmed they would experience guilt and, crucially, only if they felt responsible over the action that caused harm they would apologize.

Some previous research gives support to our hypothesis. Banakou and Slater [27] found that participants in virtual reality illusorily attributed speaking performed by their avatar to themselves, but only when they felt body ownership over that avatar. Additionally, Banakou and Slater [28] also demonstrated that the illusory agency effect only occurs when there is some real agency (visuomotor synchrony between the person's real movements and the movements of the avatar) and is not just the result of body ownership (e.g. that might be induced by visuotactile synchrony). Altogether, these works suggest that people might also experience guilt as a consequence of the actions of their avatars when they carry out actions that were not actually carried out by themselves.

To examine our hypothesis, we recruited 64 participants (42 females and 22 males) in a 2×2 between-groups factorial design. The first factor was Embodiment and the two levels were Asynchronous (the robot did not respond to the movements of the participant except for speaking), or Synchronous (where the movements of the participant were tracked and partially replicated in real-time on the robot body). The second factor was Words, where either Neutral or Offensive (insulting) words were added into the conversational stream by the robot, as if spoken by the participant. The design is shown in Table 1, and further explained in Methods.

--Table 1 about here--

## 2. Method

### 2.1. Participants

Sixty-four volunteers (42 females and 22 males), aged 18-47 (M = 23.9, SD = 5.67) took part in the experiment. Participants were recruited through ads on the university campus where the experiment took place. Most participants were university students. The volunteers received 12 Euros for their participation. Participants were naive to the purpose of the experiment. All participants gave their written informed consent prior to participating. The study was conducted with ethical approval of the [hidden for peer review].

### 2.2. System for humanoid robot embodiment

A humanoid robot unit was used for the experience of robotic embodiment. In order to create the illusion of embodiment, the participants were provided with an Oculus Rift head-mounted display (HMD) that displayed stereoscopic 3D video feed from two Microsoft HD-3000 webcams separated by a standard interocular distance at the robot's forehead. The participants were able to see the robot's body from a first person perspective when they looked down. Also, they saw the robot's body if they looked forward, towards a mirror that we placed in front of the robot (Figure 1).

The participants were also provided with a pair of headphones and a microphone so that they could communicate with the confederate. The audio was captured by a headset with an in-built microphone, located next to the robot, and the participant's words were reproduced through a pair of speakers also located next to the robot. This system allowed the participants to communicate with the confederate, who was in a different room to the participant and located in the same space as the robot during the conversation. However, the speakers and headphones were hidden to create the illusion to the participants that they talked through the robot.

In addition, the participants' arms and head in the synchronous condition were tracked and reproduced by the robot in real-time so that they had control of the robot´s body. The head was tracked with the tracker of the Oculus Rift, while the body was tracked using the Optitrack Motive Motion Capture system (Figure 1).

The robot that we used for the experiment was a Robothespian unit, manufactured by Engineered Arts, UK. This is a 180 cm. tall humanoid robot, with two legs, a torso, two arms, and a head. The joints of the robot's upper limbs are pneumatic, while the torso and head, each with three degrees of freedom, move with a DC motor. The shoulders have three degrees of freedom, the elbows and the wrist have one degree of freedom each, and the forearm has the ability to rotate along its own axis as well. The lower half of the robot was fixed in place, thus, the robot could not walk.

In the asynchronous condition, the setup was the same except for the fact that we did not provide head and body tracking so that the robot stayed still throughout the experiment. However, the participants were still able to communicate through it. Also, they saw the mirror in front of them, with the robot reflected on it.

*--Figure 1(left & right) about here--*

*Figure 1. Experimental setup. Participants (left) wear a head-mounted display (HMD), a body-tracking suit, and a pair of headphones with a microphone. In the synchronous conditions, their movements are tracked and reproduced to the robot (right). A mirror is placed in front of the robot so that the participant can identify with its body.*

## 3. Procedure

**Session 1.** After reading and signing the consent form, participants (n=64) were seated in front of a PC and were instructed to read out in a clear voice a sequence of fifty words displayed in alphabetical order. They were told that the purpose of doing the recording was to modulate their voice so that they could verbally interact using the robot on the day of the experiment. In fact, the list contained both the neutral and the offensive words in-between other words that would be later extracted and mixed with the participant dialogue in the second session. All participants recorded the same words regardless the experimental condition in which they were assigned. They came individually to the experiment and were randomly assigned to one of the conditions. After the participants left, either the neutral or the offensive words (depending on the condition) were extracted from the recording and saved as independent voice files one by one. In total, twelve words were extracted (SI, *List of words added in the conversation*).

**Session 2.** When the participants arrived at the lab, they were fitted with a body-tracking suit, a HMD and a pair of headphones with a microphone. Participants wore the HMD which displayed real-time stereoscopic 3D video feedback from the two cameras located on the robot's forehead, separated by the standard interocular distance. In the synchronous condition, participants' head movement was tracked and synchronized to the robot's head movement. Also, they were given arm movement synchronization (see *System for humanoid robot embodiment* in *Methods*). The participants followed pre-recorded instructions that were played on the headphones. In order to create the illusion of embodiment in the robot's body, they were first required to perform some simple exercises consisting of head and arm movements (SI, Fig. S1). The exercises lasted for approximately four minutes. There was a mirror in front of the robot so that participants were able to see these movements reflected on the robot (SI, Fig. S1). Also, if they looked down, they were able to see the robot's arms reproducing their own movements. In the asynchronous condition, the participants reproduced the same movements and followed the same instructions. However, the robot stayed still looking at the mirror and did not move the head or the arms.

After the embodiment part, participants were told that another participant would come and that they would have the opportunity to have a conversation with him or her through the robot. The other participant was actually a female confederate that pretended to be a participant. The confederate waited in a room annexed to the room where the robot was. She was able to hear the unfolding of the session through the speakers next to the robot, who played the instructions. When the instructions explained about the conversation with another participant, she moved to the room where the robot was. She placed herself in front of the robot, next to the mirror, so that she became visible to the participant (Fig. S1). The instructions told them that in the first part of the conversation, the other participant (i.e. the confederate, from here onwards) would be the one asking questions. After five minutes, the real participant would be the one asking questions. Once the confederate was in front of the robot, she asked the participant several general questions (e.g. work, study, hobbies). After five minutes, a pre-recorded bell sounded and it was the time for the real participant to ask questions to the confederate. After the participant had asked two or three questions the first word was added right after one of the questions. The remaining words were successively added after some of the questions so that twelve words in total sounded out loud after the participants' questions, which were distributed in the five-minute period. The first two words were always the same neutral words (i.e. *recently* and *normally*), regardless of the condition. The ten following words were either neutral or offensive words (SI, *List of words added in the conversation*), and always sounded in the same order. The participant was able to hear the words through the headphones. The words were played in the speakers next to the robot, the same ones through which the participant was talking to the confederate. The confederate was instructed to show a slightly puzzled face after the words sounded. Also, she was instructed to continue with the conversation normally, only showing a progressive mild decrease in enthusiasm so that her reaction was ambiguous and could be interpreted by the participant in a way that made sense regardless of the condition (e.g. she is offended, she is slightly bored…). After the second part of the conversation finished, the audio instructions indicated that the task had ended. Then, the instructions asked the participants whether they were willing to say something else to the confederate. If they said no, the experiment finished. If they said yes, the participants were indicated that they were free to talk. The participants' responses were recorded as part of their behavioral response for the apologies measure. After that, participants were removed from the embodiment system. Following this, the researcher told the participants that the experiment had finished and that they would meet the confederate face to face, who was in the next room. The researcher opened the door and asked the participant to follow her. Then, the researcher said hello to the confederate and stood between the participant and the confederate. The confederate also said hello to her and to the participant. Then, they both waited for the participants' comments, which were again recorded as part of their behavioral response for the apologies measure. After the participant stopped talking, the researcher indicated to the confederate that she could move to a different room to complete her part and entered back to the previous room with the real participant. Finally, the participants completed a questionnaire following which they were disclosed about the real role of the confederate and the purpose of the experiment. They were then paid and thanked for their participation and left the lab. After each session, we verified with the confederate that she did not know any of the participants.

Movie S1 contains a summary of the procedure.

**3.1. Response Variables** We had self-reported response variables for Guilt (G) and Shame (S) and behavioral responses for explicit apologies (Ext), and verbal repair (Res). See also *Measures, and Response Variables* and *Table S1 and S2* in the *SI*.

### 3.2. Measures

**Ownership and agency.** A questionnaire was adapted from previous studies [29] to measure ownership and agency. Each item (Table 2) was rated on a 7-point scale that ranged from (-3) *not at all* to (+3) *very strongly*. Reliability for the scale was α=.96.

---Table 2 about here---

**Guilt and shame.** The harder personal feelings questionnaire (PFQ2) [30] was adapted to the situation of the experiment. This questionnaire consists of 22 items and it was initially designed to measure proneness to shame (score of 0 to 40 where 40 means greatest feeling of shame) and guilt (score of 0 to 24 where 24 means greatest feeling of guilt). We adapted the questionnaire to evaluate state guilt and shame in response to the experimental manipulation (i.e. *When the robot added words during the conversation, you felt...).* The items in the questionnaire that measure guilt are *mild guilt, worry about hurting or injuring the other participant, intense guilt, regret, feeling you deserve criticism for what you did,* and r*emorse.* The items that measure shame are *embarrassment, feeling ridiculous, self-consciousness, feeling humiliated, feeling "stupid", feeling “childish”, feeling helpless or paralyzed, feelings of blushing, feeling laughable,* and *feeling disgusting to the other*. The scale also contains six control items. Reliability for the guilt subscale was α=.94 and for the shame subscale was also α=.94.

**Explicit Apologies.** We define explicit apologies in the context of the experiment as a verbal expression of regret in which the participant said the words “I am sorry” or ones with an equivalent meaning.  A binary assessment of apologies (0 is no apology, 1 is apology) was obtained based on whether the participant explicitly verbally apologized (i.e. said sorry) to the confederate or not after the conversation (see transcripts in the SI). We repeated this measure twice: right after the conversation, when the participants were asked whether they were willing to say something else to the confederate and were still embodied in the robot and after the participants were removed from the embodiment system and introduced to the confederate face to face.

**Verbal repair.** A binary assessment of verbal repair (0 is no repair, 1 is repair) was obtained based on whether the participant took verbal action to clarify the situation so that the confederate did not feel she was insulted during the conversation (see transcripts in the

SI). We examined responses to this measure twice: right after the conversation, when the participants were asked whether they were willing to say something else to the confederate and were still embodied in the robot and after the participants were removed from the embodiment system and introduced to the confederate face to face.

**Demographic measures.** Participants completed information about age and gender.

**Manipulation check.** Participants were asked whether the robot included words that they did not say during the conversation. All participants passed the manipulation check question for all conditions.

## 4. Statistical Model

--Figure 2 around here--

*Figure 2. Statistical Model. Ownership (O) is a latent variable based on the four questionnaire responses on body ownership and agency. The model is based on the idea that Ownership and type of words will influence the feelings of guilt and shame, which in turn will influence whether or not participants apologize to the confederate (explicit apologies and verbal repair).*

The model illustrated above is that the levels of embodiment (Asynchronous, Synchronous) should influence the subjective illusion of body ownership and agency, as elicited through the four questionnaire responses (myarms, mirror, agency, mybody). These four variables are posited to be different manifestations of an underlying construct that we refer to as Ownership, which is a latent variable. In turn the level of Ownership and type of words (neutral, offensive) will influence the feelings of guilt and shame. Finally, the level of guilt and shame will influence whether or not the participant apologized to the confederate (explicit apologies and verbal repair).

We explore this using a Bayesian model.

$R_i \sim ordered\_logistic(\beta_{Ro}+\beta_{R1}O_i)$ (myarms)

$M_i \sim ordered\_logistic(\beta_{Mo}+\beta_{M1}O_i)$ (mirror)

$A_i \sim ordered\_logistic(\beta_{Ao}+\beta_{A1}O_i)$ (agency)

$B_i \sim ordered\_logistic(\beta_{Bo}+\beta_{B1}O_i)$ (mybody)

$i = 1, \dots, n = 64$

Since *R* (myarms), *M* (mirror), *A* (agency) and *B* (mybody) are ordered response variables (transformed to 1,…,7 for the purposes of analysis) we use the ordered logistic model to represent them and each is a stochastically dependent on a linear function of the unknown latent variable O(wnerhip), where

$$O_i \sim N(0,10)$$

which is a prior normal distribution with mean 0 and standard deviation 10 (giving an effective possible range of -30 to 30).

$$S_i \sim N(\beta_{S0} + \beta_{S0} O_i + \beta_{S1} W_i, \sigma_s) \qquad \text{(shame)}$$

$$G_i \sim N(\beta_{G0} + \beta_{G0} O_i + \beta_{G1} W_i, \sigma_G) \qquad \text{(guilt)}$$

Here shame and guilt are normally distributed with mean depending on Ownership and Words, and the standard deviations as shown.

Finally, explicit apologies (*Ext*) and verbal repair (*Res*) are binary variables (0 = no, 1 = yes) modelled as a bernoulli_logit depending on Ownership, guilt and shame.

$$Ext_i \sim bernoulli\_logit(\beta_{Ext,0} + \beta_{Ext,1} O_i + \beta_{Ext,2} G_i + \beta_{Ext,3} S_i, \sigma_{Ext}) \qquad \text{(explicit apol.)}$$

$$Res_i \sim bernoulli\_logit(\beta_{Res,0} + \beta_{Res,1} O_i + \beta_{Res,2} G_i + \beta_{Res,3} S_i, \sigma_{Res}) \qquad \text{(verbal repair)}$$

Prior distributions:

All $\beta_{ij}$ are modelled with prior distributions N(0,10) and all $\sigma_i$ as half Cauchy distributions with scale factor 5 on the interval $(0, \infty)$.

The model was executed in Stan with 4000 iterations and 4 chains. Convergence was achieved indicated by all Rhat = 1.

Of particular interest are the posterior probabilities that the coefficients $\beta_{ij}$ (j > 0, i.e., not the intercept terms) are strictly positive (or negative). For example, if $\beta_{Ext,2} > 0$ then this indicates that guilt (*G*) is positively associated with the Explicit apologies.

Since this is not hypothesis testing, but an exploratory rather than a confirmatory study, we do not here make 'decisions' (as with significance testing) but rather report the results as posterior probabilities and 95% credible intervals. Henceforth P($\beta_{ij}$> 0), for example, refers to the *posterior probability* that $\beta_{ij} > 0$. Also of interest are the posterior standard deviation distributions. The prior distributions for the standard deviations have infinite support and their means are infinite (this is a property of the Cauchy priors).

Note that the results are not sensitive to different parameters for the priors, and following the recommendation [31] we do not use extremely noninformative prior distributions.

## 5. Results

Table 3 gives the summary statistics of the posterior distributions of the model (Note that this is one overall model, not a series of separate models).

*-- Table 3 about here--*

The latent variable Ownership is very strongly and positively related with the 4 questionnaire variables (mirror, agency, mybody, myarms). Although we did not include Embodiment as an explicit variable in the model we can see how the latent variable Ownership is related to Embodiment. See Fig. S2 in the SI, which shows the vast difference between the values of Ownership for Asynchronous compared to Synchronous (Cohen's d = 2.67). See also Fig. S3 in the SI with the scatter diagrams of Ownership by each of the questionnaire scores which illustrates both the range and scale of Ownership, and the relationship with the component questionnaire scores.

Table 3 shows the summaries of the posterior distributions of the model parameters. It is clear that the latent variable Ownership is strongly related to the individual questionnaire responses as illustrated by Fig. S3.

The following highly probable findings can be inferred from Table 3:

- The feeling of shame is positively associated with offensive words (Prob ~ 1.000).
- The feeling of guilt is positively associated with offensive words (Prob ~ 1.000).
- Explicit Apologies is positively related with guilt (Prob = 0.99)
- Verbal Repair is positively related with guilt (Prob = 0.998)

The following moderate results can be inferred:

- Ownership is associated with a *reduction* in shame (Prob = 1 – 0.104 = 0.896)
- Ownership is associated with a *reduction* in guilt (Prob = 1 – 0.165 = 0.835).
- Ownership is positively associated with explicit apologies (Prob = 0.828)
- Ownership is positively associated with verbal repair (Prob = 0.865).

There is no evidence that verbal repair is associated with shame (prob ~ 0.5, which is same as the prior probability).

Additional results and figures are reported in the SI. Additional demographic information is reported in Tables S3 and S4 (SI). Means for guilt and shame by experimental condition are represented in Table S5 (SI). The number of participants that verbalized explicit apologies and/or took verbal action to amend the situation (verbal repair) are reported in Table S6 (SI). Ownership and agency questionnaire responses by the factors Words and Embodiment are represented in Figure S4 (SI). Results for shame and guilt by the factor Embodiment are reported in Figures S5 and S6 (SI). Examples of the prior and posterior distributions of the parameter values are reported in Figures S7 and S8. Transcripts of the participants' sentences containing verbal repair or apologies are also included in the SI.

## 6. Discussion

This work examines the experience of guilt over the offensive actions of a robot avatar when these actions are clearly not the participant's fault. We found that guilt and shame were positively associated with offensive words and that apologizing and verbal repair were positively related to guilt. Also, body ownership was moderately associated to apologizing and verbal repair. Participants in experimental conditions in which the robot avatar verbally offended a confederate felt more guilt and shame and apologized more than participants assigned to control conditions (Tables S5 & S6, SI). Following previous work [6]–[12], [14], [17], the current study also shows that people are able to experience sense of embodiment of non-human but humanoid robots. This observation is also in line with studies that demonstrate that humans are able to embody avatar bodies that depart from the humanoid form in virtual reality [32]–[35].

The results thus suggest that participants felt guilty over the situation. Further studies are necessary to determine if the sense of guilt experienced by participants was directly connected to the bad actions of the robot avatar or to the participants' lack of intervention. Also, while the results clearly suggest that participants experienced guilt over the bad actions of their robot avatar, the effect of ownership is not settled. We initially expected that ownership would lead to more guilt and apologies. However, this hypothesis is only partially supported by the positive relationship between ownership and *explicit apologies* and the positive relationship between *ownership* and *verbal repair*. It seems to conflict with other moderate results suggesting a negative relationship between ownership and guilt. However, a closer look at the results of the guilt questionnaire reveals that the mean in the embodiment offensive words condition and the mean in the non-embodiment offensive words condition are practically the same (Table S3, SI). Thus, we do not find this negative relationship relevant and, in any case, we conclude that the effect of ownership was not as important in contributing to guilt as initially expected. Further work is needed to clarify the role of ownership in this regard.

The most outstanding results in terms of measures are the apologies to the confederate. These apologies clearly indicate that participants felt guilty for the misbehavior of their robot avatar and reflect their need to repair damage after they believed they caused harm to the confederate.

An important aspect of the analysis of the explicit apologies is that we only counted the result as an apology when the participant directly apologized with the intention of alleviating the suffering of the confederate. We did not count it as an apology when the participants brought up the issue just to avoid shame. There was a participant in the sync. neutral words condition who apologized to the confederate. After concluding the experiment, she explained that the reason for apologizing was because she considered she was responsible for the bad quality of the interaction when the neutral words appeared to be nonsense in the middle of the conversation.

Regarding shame, we measured this emotion as a complementary emotion to guilt in order to clearly distinguish one from the other. There is no evidence that verbal repair is associated with shame, which supports the theoretical approach that we adopted.

An interesting aspect to report is that most participants waited for the conversation to finish before mentioning about the words. Almost no one interrupted the conversation and only clarified the situation when they believed the experiment had finished. When participants were asked about the reasons why they did not interrupt the conversation to clarify that they were not saying the words, most of them reported that despite experiencing a very uncomfortable feeling because of the situation they did not want to ruin the experiment so they decided to wait until the end.

As a cautionary note, our study included a relatively small sample size, with over representation of women as participants and an unequal distribution by gender across experimental conditions. Although our study has investigated a quite new area of research, and we have found some strong conclusions in terms of probability, this type of study requires further work in order to understand better what might go wrong and the human response to it in these types of robotically embodied interactions.

To conclude, embodiment systems in virtual reality and robots are progressively expanding in the society [36]. Thus, there is an increasing need to examine the moral, ethical, and legal aspects of these technologies [37] as well as the behavioral effects [38]–[41] that their uses can entail. Our work specifically explores the moral emotions and behaviors in this context and reveals that participants may integrate a robot avatar as part of themselves and feel guilty for their actions. Addressing the moral emotions and behaviors resulting from interactions involving physical and digital avatars is crucial to facilitate the regulation of embodiment technologies in the society.

## 7. Competing Interests

No competing financial interests exist.

## 8. References

[1] W. Steptoe, J. Normand, S. Superiore, S. Anna, A. Steed, J. Kautz, and M. Slater,

"Acting Rehearsal in Collaborative Multimodal Mixed Reality," *Presence Teleoperators Virtual Environ.*, vol. 21, no. 4, pp. 406–422, 2012.

[2] D. Perez-Marcos, M. Solazzi, W. Steptoe, O. Oyekoya, A. Frisoli, T. Weyrich, A. Steed, F. Tecchia, M. Slater, and M. V. Sanchez-Vives, "A fully immersive set-up for remote interaction and neurorehabilitation based on virtual body ownership," *Front. Neurol.*, vol. JUL, 2012.

[3] S. Kishore, X. Navarro, E. Dominguez, N. de la Peña, and M. Slater, "Beaming into the News: A System for and Case Study of Tele-Immersive Journalism," *IEEE Comput. Graph. Appl.*, p. In Press, 2016.

[4] A. Steed, W. Steptoe, W. Oyekoya, F. Pece, T. Weyrich, J. Kautz, D. Friedman, A. Peer, M. Solazzi, F. Tecchia, M. Bergamasco, and M. Slater, "Beaming: An asymmetric telepresence system," *IEEE Comput. Graph. Appl.*, vol. 32, no. 6, pp. 10–17, 2012.

[5] S. Kishore, X. Navarro Muncunill, P. Bourdin, K. Or-Berkers, D. Friedman, M. Slater, X. N. Muncunill, P. Bourdin, K. Or-Berkers, D. Friedman, and M. Slater, "Multi-Destination Beaming: Apparently Being in Three Places at Once through Robotic and Virtual Embodiment," *Front. Robot. AI*, vol. 3, no. November, p. 65, 2016.

[6] M. Alimardani, S. Nishio, and H. Ishiguro, "Humanlike robot hands controlled by brain activity arouse illusion of ownership in operators.," *Sci. Rep.*, vol. 3, p. 2396, Jan. 2013.

[7] L. Aymerich-Franch, D. Petit, G. Ganesh, and A. Kheddar, “Embodiment of a humanoid robot is preserved during partial and delayed control,” in *2015 IEEE International Workshop on Advanced Robotics and its Social Impacts*, 2015.

[8] L. Aymerich-Franch, D. Petit, G. Ganesh, and A. Kheddar, “Non-human Looking Robot Arms Induce Illusion of Embodiment,” *Int. J. Soc. Robot.*, vol. 9, no. 4, pp. 479–490, 2017.

[9] L. Aymerich-Franch, D. Petit, G. Ganesh, and A. Kheddar, “Object Touch by a Humanoid Robot Avatar Induces Haptic Sensation in the Real Hand,” *J. Comput. Commun.*, vol. 22, no. 4, pp. 215–230, 2017.

[10] L. Aymerich-Franch, D. Petit, G. Ganesh, and A. Kheddar, “The second me: Seeing the real body during humanoid robot embodiment produces an illusion of bi-location,” *Conscious. Cogn.*, vol. 46, pp. 99–109, 2016.

[11] O. Cohen, S. Druon, S. Lengagne, A. Mendelsohn, R. Malach, A. Kheddar, and D. Friedman, “fMRI robotic embodiment: A pilot study,” *2012 4th IEEE RAS EMBS Int. Conf. Biomed. Robot. Biomechatronics*, pp. 314–319, Jun. 2012.

[12] O. Cohen, M. Koppel, R. Malach, and D. Friedman, “Controlling an avatar by thought using real-time fMRI,” *J. Neural Eng.*, vol. 11, no. 3, 2014.

[13] S. Kishore, M. González-Franco, C. Hintemüller, C. Kapeller, C. Guger, M. Slater, and K. J. Blom, “Comparison of SSVEP BCI and Eye Tracking for Controlling a Humanoid Robot in a Social Environment,” *Presence Teleoperators Virtual Environ.*, vol. 23, no.

3, pp. 242–252, 2014.

[14] S. Kishore, X. Navarro, E. Dominguez, N. de la Peña, and M. Slater, "Beaming into the News: A System for and Case Study of Tele-Immersive Journalism," *IEEE Comput. Graph. Appl.*, p. In Press, 2016.

[15] M. Lombard and T. Ditton, "At the Heart of it All. The Concept of Presence," *J. Comput. Commun.*, vol. 3, no. September, p. 20, 1997.

[16] T. Metzinger, "Two principles for robot ethics," in *Robotik und Gesetzgebung*, E. Hilgendorf and J. P. Günther, Eds. Baden-Baden: Nomos, 2013, pp. 263–302.

[17] S. Kishore, M. González-Franco, C. Hintemüller, C. Kapeller, C. Guger, M. Slater, and K. J. Blom, "Comparison of SSVEP BCI and Eye Tracking for Controlling a Humanoid Robot in a Social Environment," *Presence Teleoperators Virtual Environ.*, vol. 23, no. 3, pp. 242–252, 2014.

[18] R. F. Baumeister, A. M. Stillwell, and T. F. Heatherton, "Guilt: An interpersonal approach.," *Psychol. Bull.*, vol. 115, no. 2, pp. 243–267, 1994.

[19] I. E. de Hooge, M. Zeelenberg, and S. M. Breugelmans, "Moral sentiments and cooperation: Differential influences of shame and guilt," *Cogn. Emot.*, vol. 21, no. 5, pp. 1025–1042, 2007.

[20] I. E. de Hooge, R. M. A. Nelissen, S. M. Breugelmans, and M. Zeelenberg, "What Is Moral About Guilt? Acting ' Prosocially' at the Disadvantage of Others," *J. Pers. Soc. Psychol.*, vol. 100, no. 3, pp. 462–473, 2011.

[21] J. P. Tangney, R. S. Miller, L. Flicker, and D. H. Barlow, “Are shame, guilt, and embarrassment distinct emotions?,” *J. Pers. Soc. Psychol.*, vol. 70, no. 6, pp. 1256–1269, 1996.

[22] J. P. Tangney, J. Stuewig, and D. J. Mashek, “Moral Emotions and Moral Behavior,” *Annu. Rev. Psychol.*, vol. 58, no. 1, pp. 345–372, 2007.

[23] D. Keltner, “Evidence for the Distinctness of Embarrassment, Shame, and Guilt: A Study of Recalled Antecedents and Facial Expressions of Emotion,” *Cogn. Emot.*, vol. 10, no. 2, pp. 155–172, 1996.

[24] J. P. Tangney, R. S. Miller, L. Flicker, and D. H. Barlow, “Are shame, guilt, and embarrassment distinct emotions?,” *J. Pers. Soc. Psychol.*, vol. 70, no. 6, pp. 1256–1269, 1996.

[25] D. Keltner, “Evidence for the Distinctness of Embarrassment, Shame, and Guilt: A Study of Recalled Antecedents and Facial Expressions of Emotion,” *Cogn. Emot.*, vol. 10, no. 2, pp. 155–172, 1996.

[26] de Hooge, I. E. de Hooge, R. M. A Nelissen, S. M. Breugelmans, and M. Zeelenberg, “What is moral about guilt? Acting &quot; prosocially &quot; at the disadvantage of others,” *J. Pers. Soc. Psychol.*, vol. 100, no. 3, pp. 462–473, 2011.

[27] D. Banakou and M. Slater, “Body ownership causes illusory self-attribution of speaking and influences subsequent real speaking,” *Proc. Natl. Acad. Sci.*, vol. 111, no. 49, pp. 17678–17683, 2014.

[28] D. Banakou and M. Slater, “Embodiment in a virtual body that speaks produces agency over the speaking but does not necessarily influence subsequent real speaking,” *Sci. Rep.*, vol. 7, no. 1, 2017.

[29] D. Banakou and M. Slater, “Body ownership causes illusory self-attribution of speaking and influences subsequent real speaking,” *Proc. Natl. Acad. Sci.*, vol. 111, no. 49, pp. 17678–17683, 2014.

[30] D. H. Harder and A. Zalrna, “Two Promising Shame and Guilt Scales: A Construct Validity Comparison,” *J. Pers. Assess.*, vol. 55, no. 3–4, pp. 729–745, 1990.

[31] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, P. Li, and A. Riddell, “Stan: A Probabilistic Programming Language,” *J. Stat. Softw.*, vol. 76, no. 1, pp. 1–32, 2017.

[32] S. J. G. Ahn, J. Bostick, E. Ogle, K. L. Nowak, K. T. McGillicuddy, and J. N. Bailenson, “Experiencing Nature: Embodying Animals in Immersive Virtual Environments Increases Inclusion of Nature in Self and Involvement With Nature,” *J. Comput. Commun.*, vol. 21, no. 6, pp. 399–419, 2016.

[33] L. Aymerich-Franch, “Can We Identify with a Block ? Identification with Non-anthropomorphic Avatars in Virtual Reality Games,” in *Proceedings of the International So-ciety for Presence Research Annual Conference.*, 2012.

[34] W. Steptoe, A. Steed, and M. Slater, “Human tails: ownership and control of extended humanoid avatars.,” *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 4, pp.

583–90, 2013.

[35] A. S. Won, J. Bailenson, J. Lee, and J. Lanier, “Homuncular Flexibility in Virtual Reality,” *J. Comput. Commun.*, vol. 1999, p. n/a-n/a, 2015.

[36] F. Biocca, “The Cyborg ’ s Dilemma : Progressive Embodiment in Virtual Environments Minding the Body , the Primordial Communication Medium,” *JCMC*, vol. 3, no. September, pp. 1–29, 1997.

[37] K. Gabriels, K. Poels, and J. Braeckman, “Morality and involvement in social virtual worlds: The intensity of moral emotions in response to virtual versus real life cheating,” *New Media Soc.*, vol. 16, no. 3, pp. 451–469, 2014.

[38] L. Aymerich-Franch, R. F. Kizilcec, and J. N. Bailenson, “The Relationship between Virtual Self Similarity and Social Anxiety,” *Front. Hum. Neurosci.*, vol. 8, no. November, pp. 1–10, 2014.

[39] V. Groom, J. N. Bailenson, and C. Nass, “The influence of racial embodiment on racial bias in immersive virtual environments,” *Soc. Influ.*, vol. 4, no. 3, pp. 231–248, Jul. 2009.

[40] R. S. Rosenberg, S. L. Baughman, and J. N. Bailenson, “Virtual Superheroes: Using Superpowers in Virtual Reality to Encourage Prosocial Behavior,” *PLoS One*, vol. 8, no. 1, pp. 1–9, 2013.

[41] N. Yee and J. Bailenson, “The proteus effect: The effect of transformed self-representation on behavior,” *Hum. Commun. Res.*, vol. 33, no. 3, pp. 271–290,

2007.

## Tables

Table 1

*Experimental design. A 2× 2 between groups design with two binary factors: Embodiment and Words. N = 16 participants in each cell.*

| Embodiment(E)/Words(W) | Neutral Words | Offensive Words |
|---|---|---|
| Asynchronous | 16 | 16 |
| Synchronous | 16 | 16 |

Table 2

*Body ownership and agency questions. Each question was scored on a -3 (disagree) to 3 (agree) scale. The variable names and abbreviations in the first two columns are used in the statistical model.*

| Variable Name | Abbreviation | Question |
|---|---|---|
| myarms | R | I felt as if the hands of the robot were my hands, even though they did not look like me |
| mirror | M | I felt as if the body I saw in the mirror was my body, even though it did not look like me |
| agency | A | I felt as if I could control the robot's body as if it was mine |
| mybody | B | I felt as if the body I saw was my body, even though it did not look like me |

Table 3

*Summary statistics of the posterior distributions of the model*

| | | Mean and Standard Deviation of the parameter posterior distribution | | 95% credible interval of the posterior distribution | | |
|---|---|---|---|---|---|---|
| Parameter | Coefficient of.. | Mean | SD | 2.5 percentile | 97.5 percentile | P( > 0) |
| $\beta_{mirror,0}$ | | 0.04 | 10.175 | -19.96 | 20.08 | 0.503 |
| $\beta_{mirror,1}$ | Ownership | 0.56 | 0.117 | 0.36 | 0.82 | 1.000 |
| $\beta_{agency,0}$ | | 0.05 | 9.802 | -19.11 | 19.04 | 0.502 |
| $\beta_{agency,1}$ | Ownership | 0.40 | 0.083 | 0.26 | 0.59 | 1.000 |
| $\beta_{mybody,0}$ | | 0.01 | 9.944 | -19.58 | 19.36 | 0.507 |
| $\beta_{mybody,1}$ | Ownership | 0.57 | 0.115 | 0.37 | 0.82 | 1.000 |
| $\beta_{myarms,0}$ | | 0.05 | 9.968 | -19.52 | 19.78 | 0.501 |
| $\beta_{myarms,1}$ | Ownership | 25.36 | 6.930 | 12.97 | 40.26 | 1.000 |
| $\beta_{shame,0}$ | | 10.12 | 1.601 | 6.95 | 13.24 | 1.000 |
| $\beta_{shame,1}$ | Ownership | -0.16 | 0.127 | -0.41 | 0.09 | 0.104 |
| $\beta_{shame,2}$ | Words | 11.21 | 2.217 | 6.88 | 15.67 | 1.000 |
| $\sigma_{shame}$ | | 9.21 | 0.833 | 7.74 | 11.06 | 1.000 |
| $\beta_{guilt,0}$ | | 2.59 | 0.976 | 0.65 | 4.52 | 0.997 |
| $\beta_{guilt,1}$ | Ownership | -0.07 | 0.076 | -0.22 | 0.07 | 0.165 |
| $\beta_{guilt,2}$ | Words | 10.25 | 1.368 | 7.57 | 12.95 | 1.000 |
| $\sigma_{guilt}$ | | 5.46 | 0.495 | 4.59 | 6.53 | 1.000 |
| $\beta_{Ext,0}$ | | -4.55 | 1.248 | -7.29 | -2.43 | 0.000 |
| $\beta_{Ext,1}$ | Ownership | 0.05 | 0.049 | -0.05 | 0.15 | 0.828 |
| $\beta_{Ext,2}$ | guilt | 0.29 | 0.129 | 0.05 | 0.56 | 0.990 |
| $\beta_{Ext,3}$ | shame | -0.02 | 0.086 | -0.19 | 0.15 | 0.390 |
| $\beta_{Res,0}$ | | -2.50 | 0.753 | -4.09 | -1.14 | 0.000 |
| $\beta_{Res,1}$ | Ownership | 0.05 | 0.044 | -0.04 | 0.14 | 0.865 |
| $\beta_{Res,2}$ | guilt | 0.27 | 0.105 | 0.08 | 0.49 | 0.998 |
| $\beta_{Res,3}$ | shame | 0.00 | 0.068 | -0.13 | 0.13 | 0.523 |

**Figures**

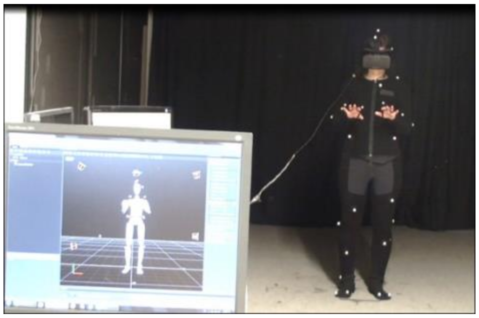


Figure 1 - left

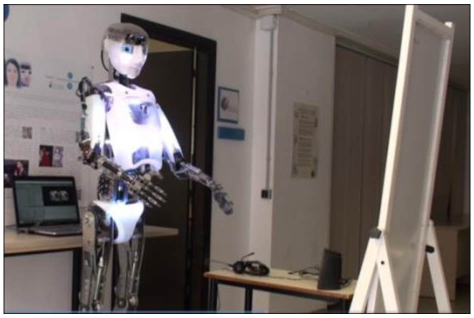


Figure 1 – right

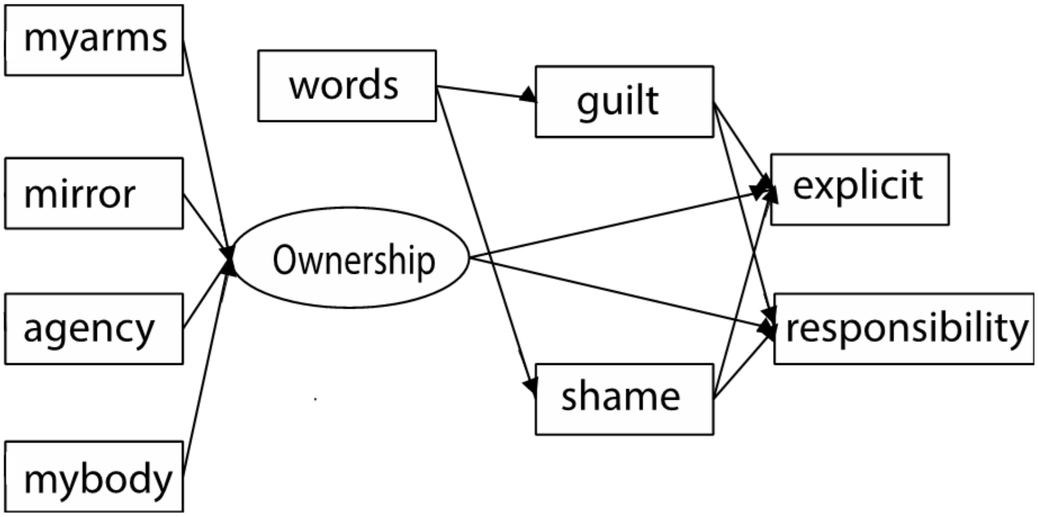


Figure 2

Supporting Information for:

When your Robot Avatar Misbehaves you are Likely to Apologize: an exploration of guilt during robot embodiment

## List of words added in the conversation

Words added in the offensive words condition: Recently, Normally, Ugly, Glasses, Loser, Bitter, Potato, Stupid, Fat, Boring, Darling, Idiot.

Words added in the neutral words condition: Recently, Normally, Currently, Lately, Maybe, Also, Soon, Similarly, Previously, Besides, Often, Always.

## Response Variables

Note that E (explicit apologies) and R (verbal repair) were coded by two independent coders.

| | **explicit2** | | |
|---|---|---|---|
| explicit1 | **0** | **1** | **Total** |
| | | | |
| 0 | 50 | 1 | 51 |
| 1 | 0 | 9 | 9 |
| | | | |
| Total | 50 | 10 | 60 |

*Table S1. Frequency of agreement for explicit apologies*

Table S1 shows the frequency of agreement. Cohen’s Kappa = 0.94 which is a very good correspondence. As will be seen there were 4 missing values from Coder 2. Therefore, only the results for Coder 1 were used.

| | **repair 2** | | |
|---|---|---|---|
| **repair1** | **0** | **1** | **Total** |
| | | | |
| **0** | 35 | 0 | 35 |

| | | | |
|---|---|---|---|
| **1** | 0 | 25 | 25 |
| | | | |
| **Total** | 35 | 25 | 60 |

*Table S2. Frequency of agreement for verbal repair*

Table S2 shows frequency of agreement for verbal repair. As can be seen, there was complete agreement between the coders, Cohen’s Kappa = 1. However, again there were 4 missing values from Coder 2, and therefore Coder 1 is used.

**Additional results**

We initially examined group distribution regarding demographic variables (age and gender). Age mean was similar across groups (Table S3). However, we had a bigger number of females than males in the sample and groups were not equally distributed regarding gender (Table S4). Since gender was not equally distributed across experimental conditions, we examined potential effects of gender if it was added in the model as an explanatory variable for guilt and shame. When this variable was added in the model, we found that gender was not implicated in body ownership. Women were less likely to feel guilt (Prob = 0.83). It did not influence shame at all. Other than the effects on guilt, gender did not influence the results at all. Thus, we did not include gender in the final model.

| | **Words** | |
|---|---|---|
| **Embodiment** | **Offensive** | **Neutral** |
| **Synchronous** | | |
| **Age** | 26.37 (1.6) | 23.75(.98) |
| | | |
| **Asynchronous** | | |
| **Age** | 21.81(.86) | 23.62(1.8) |

Table S3. *Mean (SE) for age, by experimental condition.*

| | **Words** | |
|---|---|---|

| Embodiment | Offensive | Neutral |
|---|---|---|
| Synchronous | | |
| Females | 7 | 14 |
| Males | 9 | 2 |
| | | |
| Asynchronous | | |
| Females | 9 | 12 |
| Males | 7 | 4 |

Table S4. *Gender distribution by experimental condition.*

Table S5 shows mean and SE for guilt and and shame, by experimental condition.

| | Words | |
|---|---|---|
| Embodiment | Offensive | Neutral |
| Synchronous | | |
| Guilt | 12.8 ± 1.64 | 1.7 ± 0.85 |
| Shame | 19.9 ± 2.62 | 8.4 ± 2.12 |
| | | |
| Asynchronous | | |
| Guilt | 13.1 ± 1.81 | 3.4 ± 0.87 |
| Shame | 23.3 ± 2.55 | 11.9 ± 1.83 |

Table S5. *Mean and SE for guilt and shame, by experimental condition.*

Table S6 shows number of participants that verbalized explicit apologies (i.e. said sorry or similar), and number of participants that took verbal action to amend the situation during or after the conversation (verbal repair), by experimental condition. We also analyzed perceived concern in the participants' voice just after the conversation, when they had the opportunity to talk to the confederate. We rated perceived concern on their voice on a 7-point scale (1= not at all concerned, 7 = extremely concerned). We did not include this measure in the main model because Cohen Kappa was slightly low (k = .597, p<.001). This is not surprising because we used a 7-point scale. Thus, it was unlikely that both coders would rate perceived concern exactly the same way. We report the results of perceived concern in the participants' voice (mean, SD, average of the two coders) also in Table S6.

As can be observed in Table S6, participants in the synchronous offensive words condition are the ones who apologized more, showed more verbal repair, and expressed more concern on their voices, followed by participants in the asynchronous offensive words conditions. The number of explicit apologies and verbal repair in the neutral conditions was close to zero. The participants' voice in the neutral conditions did not show any or nearly any concern.

Concerned voice significantly correlated with guilt (r=.696, n=60, p<.001) and shame (r=.626, n=60, p<.001). Explicit apologies also correlated with guilt (r=.540, n=64, p<.001) and shame (r=.441, n=64, p<.001). So did verbal repair with guilt (r=.630, n=64, p<.001) and shame (r=.543, n=64, p<.001).

| | **Words** | |
|---|---|---|
| **Embodiment** | **Offensive** | **Neutral** |
| **Synchronous** | 7 | 1 |
| **Explicit Apologies** | 13 | 2 |
| **Verbal Repair** | 3.57 (2.4) | 1.1(.2) |
| **Concerned voice** | | |
| | | |
| **Asynchronous** | | |
| **Explicit Apologies** | 3 | 0 |
| **Verbal Repair** | 10 | 1 |
| **Concerned voice** | 2.94(2.1) | 1.22(.87) |

Table S6. Number of participants that apologized by condition and M(SE) for concerned voice

## 9. Further basic graphs and supporting figures

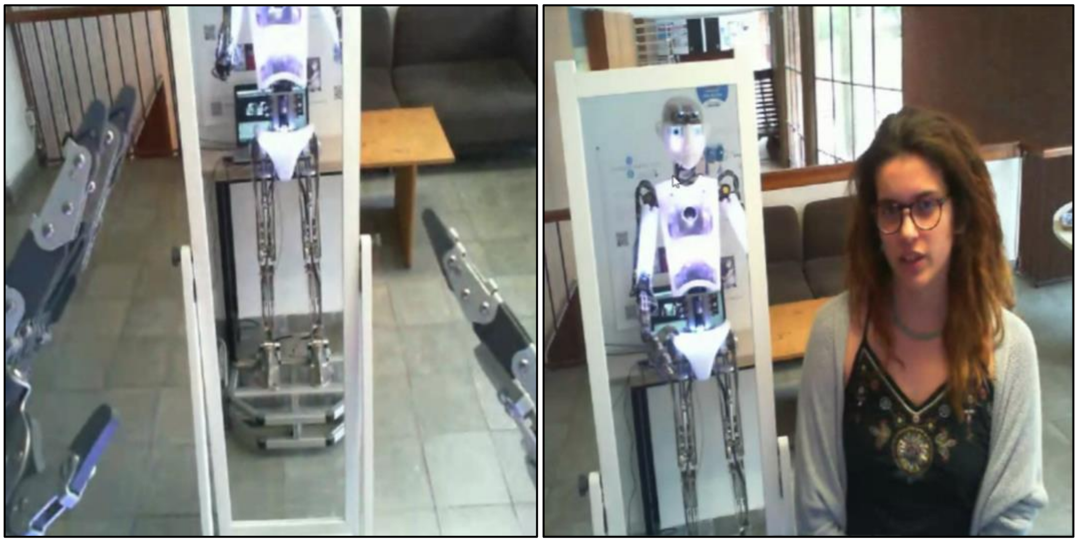


Figure S1. *Participants' field of view during the experiment. On the left, participants' view during the embodiment exercises: they were able to see the robotic limbs when they looked down. On the right, participants' view during the interaction with the confederate.*

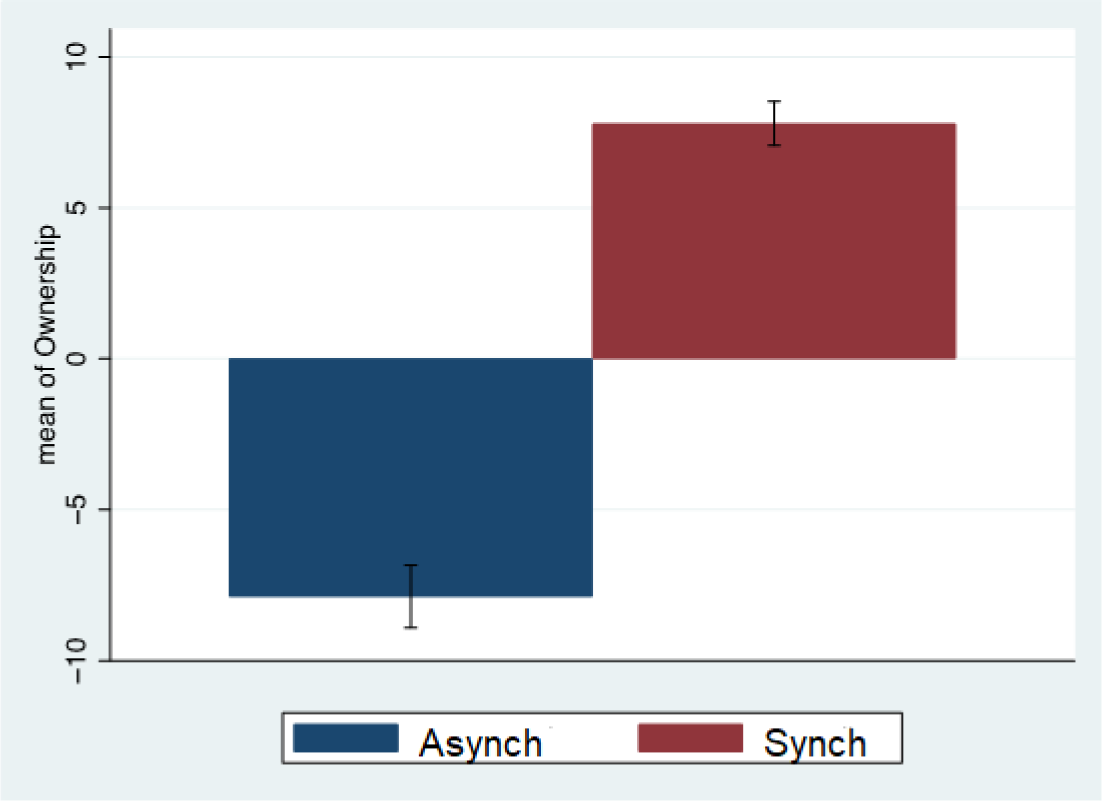


Figure S2. Means and Standard Errors of the distributions of Ownership (1…n) by Embodiment

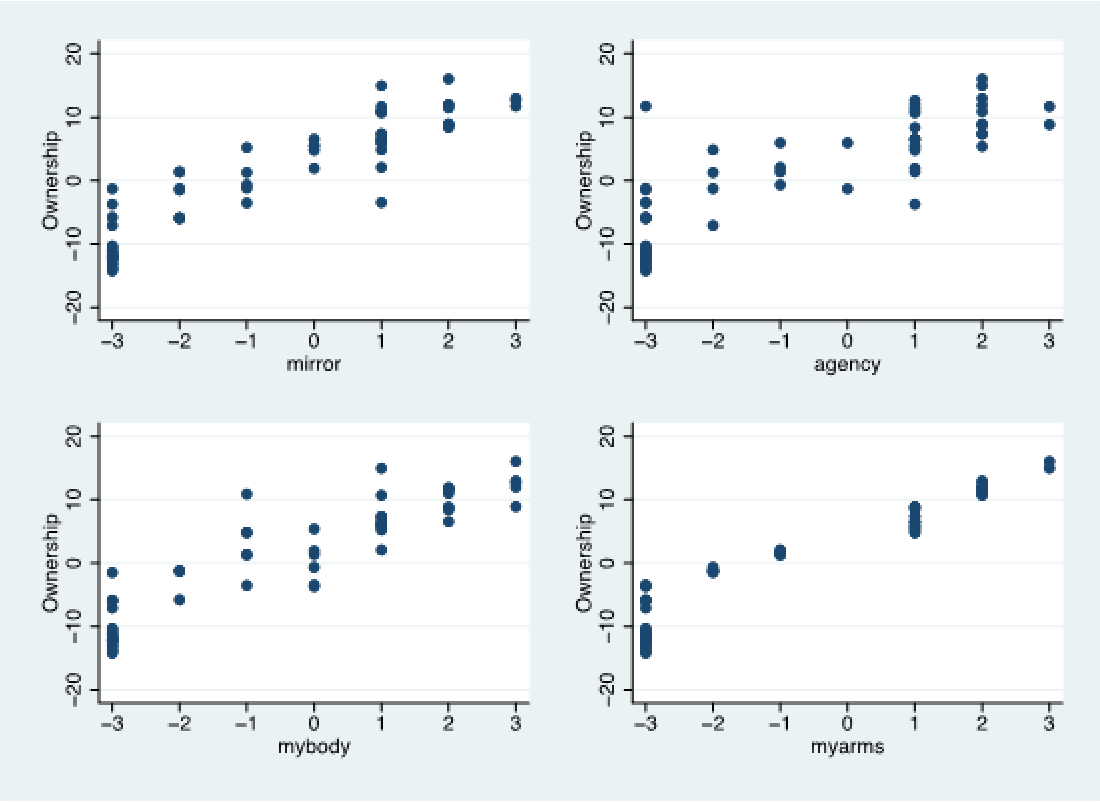


*Figure S3. Means of the distributions of Ownership (1..n) by each of the ownership and agency questionnaire scores*

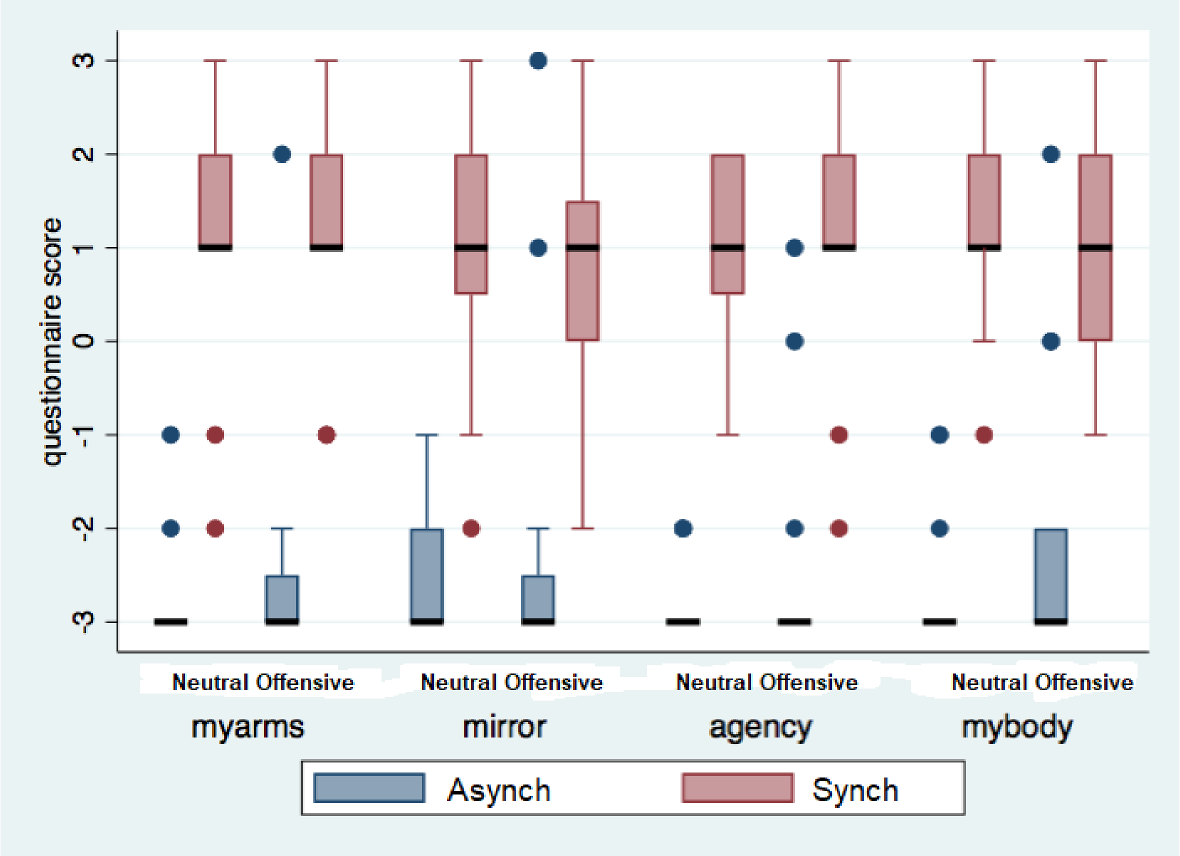
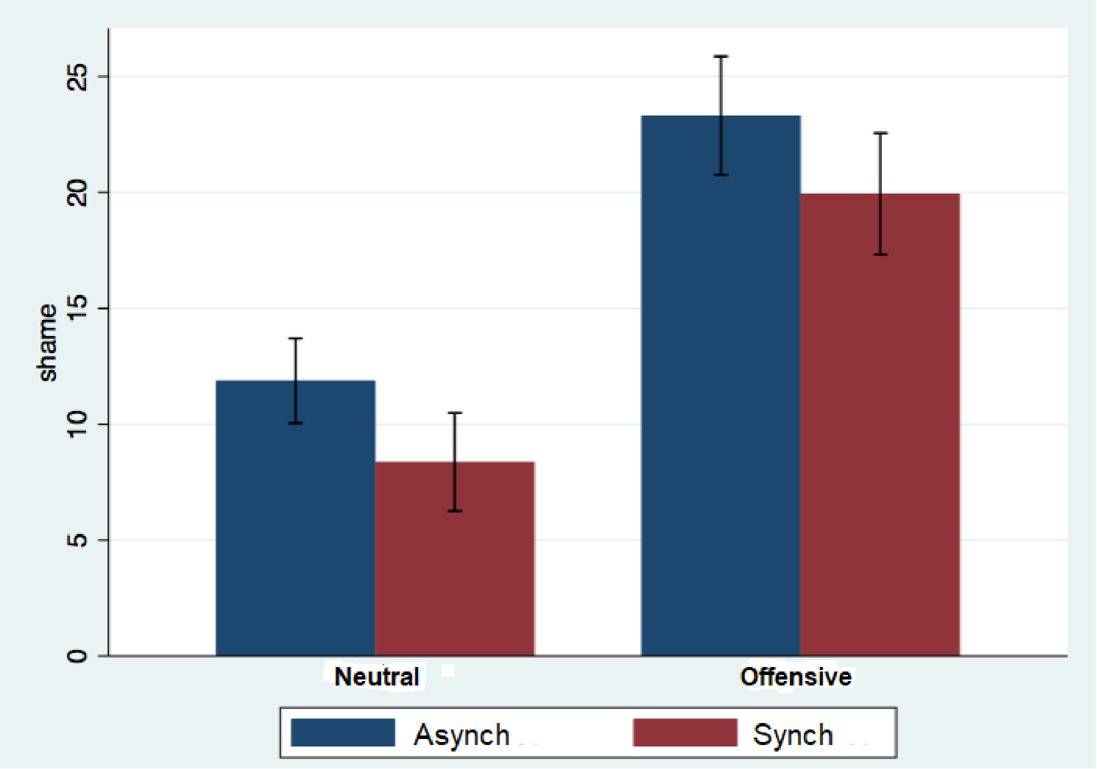


Figure S4. *Box plot of the ownership and agency questionnaire responses by the factors Words and Embodiment*

Figure S4 shows ownership and agency questionnaire responses item by item by the factors Words and Embodiment. It is clear that there is a strong effect due to Embodiment but the type of words did not have any effect on the sense of embodiment.



**10.**

Figure S5. *Bar chart (means and standard errors) for Shame*

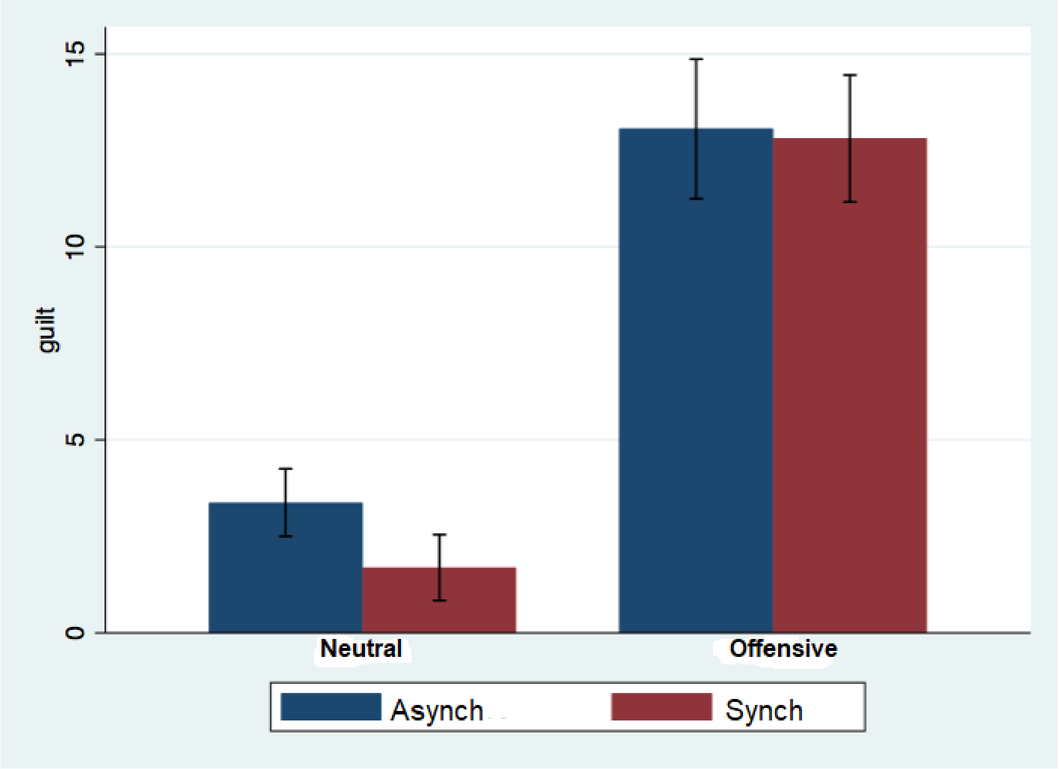


Figure S6. *Bar chart (means and standard errors) for Guilt*

Figures S5 and S6 show the results for shame and guilt by the factor Embodiment.

## 11. Prior and Posterior Distributions of the Model Parameters

Below are two examples of the prior and posterior distributions of the parameter values.

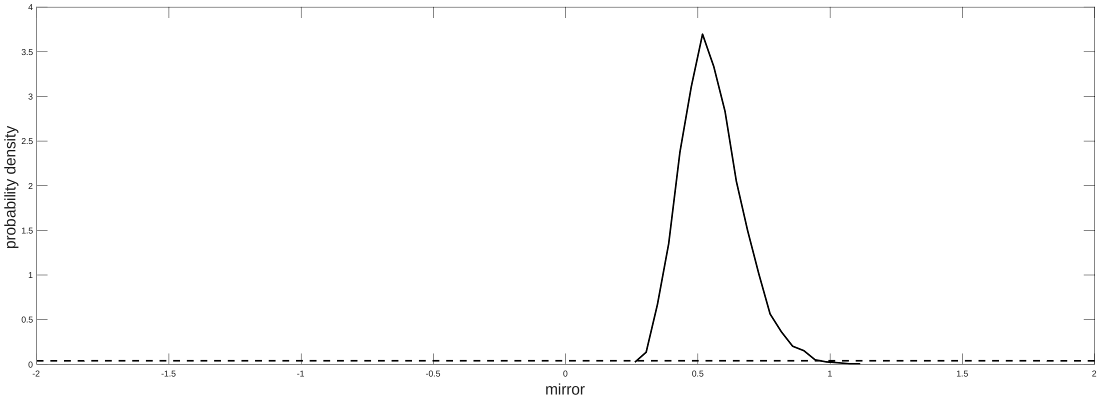


*Figure S7. Prior and Posterior distributions for $\beta_{mirror,1}$ – the dashed line is the prior distribution*

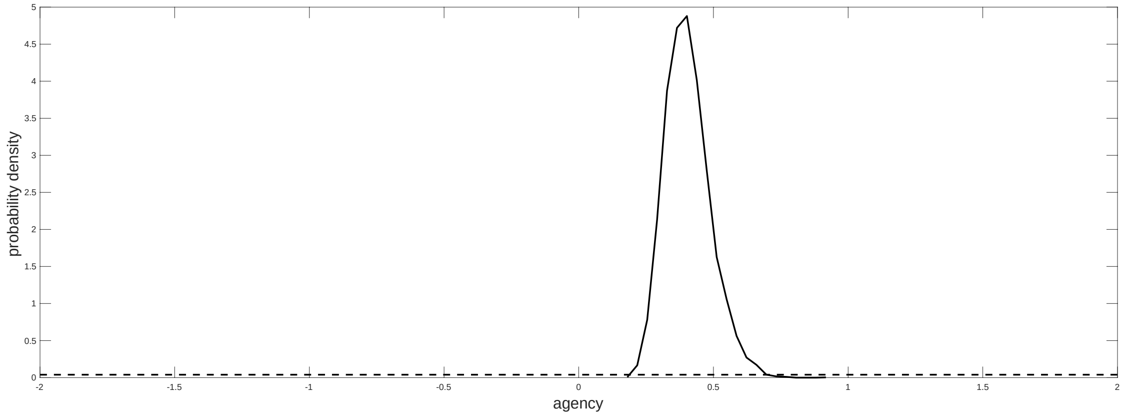


*Figure S8. Prior and Posterior distributions for $\beta_{agency,1}$ – the dashed line is the prior distribution*

**Transcript of the participants' sentences during and after the conversation aimed at clarifying the situation so that the confederate did not feel insulted, and apologies (in capital letters)**

Apologies and clarifying sentences in the synchronous, bad words condition:

- (102, T2) LO SIENTO MUCHÍSIMO (I am really sorry)
- (106, T2) I AM SO SORRY
- (107, T1) I AM SORRY for those words like stupid, fat…
- (107, T2) I AM SORRY one more time, it was not me
- (108, during the conversation) Someone is talking instead of me / There is some word interference / I didn't say that, my other personality said that / I didn't say glasses / I didn't say stupid
- (109, T1) EM SAP MOLT DE GREU (I am really sorry)
- (114, T1) todas esas palabras que iban metiendo por ahí no las decía yo, te veía la cara, y no sabía que hacer… LO SIENTO (all these words they added, I did not say them, I was seeing your face and I didn't know what to do…I am sorry)
- (114, T2) Esto es una putada, que le estabas dicendo un montón de cosas, me estaba sintiendo fatal (This sucks, I was saying her lots of things, I was feeling really bad)
- (202, T2) ¿Te decía como cosas raras del palo, de repente te llamaba fea, escuchaba yo, el robot? Hmm, pues no se, el experimento muy interesante… (Did I tell you strange things like, suddenly said you ugly, I heard that, the robot? Hmm, ok I don't know, the experiment was interesting…)
- (203, T1) no sé si tu escuchabas unas palabras que eran feas, AI QUE PENA CONTIGO, eran feas pero yo no las estaba diciendo, LAMENTO SI FUE UNA MALA INTERACCIÓN, LO SIENTO…ai no, sonaron muy feas! (I am not sure if you heard those very ugly words, I feel so bad for you, these were ugly but I was not saying them, I am so sorry if the interaction was bad, I am sorry…oh no, they really sounded ugly!)

- (203, T2) Ai QUÉ PENA CONTIGO esas palabras sonaron horribles…fuertes, muy fuertes… QUÉ PENA CONTIGO (I feel so sorry for you, these words sounded horrible…strong, so strong…I feel so sorry for you)
- (205, T1) te quería preguntar una cosa, había unas palabras que se escuchaban, tu las escuchabas? Porque esas palabras no las decía yo, eran palabras que no iban con la conversación, eran como ofensivas, por si te molestaron ME DISCULPO (I wanted to ask you something, there were some words being heard, did you hear them? Because I didn't say these words, they were words that did not go with the conversation, like offensive. In case they bothered you, I apologize)
- (210, T1) habrás notado que había unas palabras que salían como de mi propia voz, como imbécil, gorda… y yo no las estaba diciendo, directamente en este momento,entonces se me hizo un poco extraño, también porque como reaccionabas cada vez que te veía cuando se decían estas palabras, pero eran pregrabadas (you will have noticed that there were some words said like from my own voice, like stupid, fat…and I was not saying them, directly on that moment, so it was a little bit weird, also because I was seeing the way you reacted every time these words were said, but they were pre-recorded!)
- (210, T2) creo que fue un poco extraño, pero no dije nada, yo te juro que no dije nada, esas palabras eran con mi propia voz pero estaban gravadas mucho antes de este momento y no tenía ni idea de que…por eso me quedé pensando si tenía que decirte cuando estábamos conversando, además se me hizo un poco estraño porque no soy yo, o sea, no soy yo...no, o sea, es muy exraño en realidad, las palabras que salían sonaban muy fuerte pero al rato que me miraba al espejo era como, algo no cuadra, entonces, por si acaso si te sentiste un poco no se...(It was weird but I did not say anything, I swear I did not say anything, these words were with my own voice but they were recorded much earlier than this moment and I had no idea that…that's why I kept thinking if I had to tell you something while we were talking, also it felt really strange because it was not me, it's not me…no, really, it's really strange actually, the words that came sounded very strongly but at the same time I was looking myself at the mirror and it was like, something is not right, so, in case you felt a little bit…I don't know…)
- (238, T1) During the conversation: no soy yo el que dice eso! (I am not the one saying these things!)

- (241, T1) Te has sentido incómoda con las palabras que iba diciendo? Bueno no es que yo escuchaba diferente…es que te noté más incómoda ahora que cuando me hiciste las preguntas…(did you feel uncomfortable with the words I was saying? Well, maybe I heard something different…I felt you were more uncomfortable now than when you were asking me questions…)

Apologies and clarifying sentences in the asynchronous, bad words condition:

- (208, T2) Hi havia algunes paraules pregravades que no era…sonava aquella paraula per despistar, imbècil, patata... (There were some prerecorded words that were not…the word sounded to confuse…stupid, potato…)
- (214, T1) Lo de fea amargada y aburrida no lo digo yo eh, son gravaciones que pasan después (This thing about bitter, boring…I do not say them eh! They are recordings that come after…)
- (214, T2) De repente fea, gorda y yo buenooo, no las digo yo (Suddenly ugly, fat, and I likeee, I am not saying that)
- (215, T1) No et sentis ofesa per les coses que has escoltat, no les volia dir jo, o sigui, no les he dit (Do not feel offended by the things you just heard, I didn't want to say them, well, I didn't say them)
- (215, T2) Jo no estava dient ni avorrida ni lletja ni res d'aquestes coses, semblava que era jo que ho deia? Jo anava parlant i de cop sento que la meva veu mateixa diu lletja, avorrida, i jo, no no, això no ho estic dient (I was not saying boring or ugly or any of these things, did it seem like I was saying them? I was talking and suddenly I hear my own voice says ugly, boring, and I, no, no, I am not saying this)

- (218, T1) Has anat sentit paraules que no et deia jo, patata, flor…les havíem gravat prèviament i anaven sortint al mig de la conversa...però no t'he insultat (you were hearing words that I was not telling you, potato, darling…we recorded them previously and appeared at the middle of the conversation…but I did not insult you)
- (218, T2) Que no t'he insultat, que no era jo, és que anava dient coses i era com,ostres, no sabia…patata, gorda,no sé què més he dit (I didn't insult you, it wasn't me, I was saying things and it was like, damn, I didn't know…potato, fat…I don't know what else I said)

- (219, T2) Hay algunas cosas que te han dicho que yo no te las he dicho eh…ahora entenderás un poco más (there were some things they said to you that I didn't say to you ah…you will understand a little bit more now…)
- (225, during the conversation): això no ho he dit jo / alaaa / sentiràs paraules que jo no dic, no t'asustis si us plau / ualaaa / uala / òstres (I didn't say that / wooow / you will hear words that I am not saying do not be afraid please / woow / woow / jeez)
- (231, T1) Las palabras esas que han sonado idiota, imbécil, no las estaba diciendo yo (the words that sounded, idiot, stupid…I wasn't saying them)
- (234, T2) Yo no te insultaba eh, era parte del experimento, gravé las palabras hace una semana (I wasn't insulting eh, it was part of the experiment, I recorded the words a week ago)
- (238, T2) Se siente rarísimo, sobretodo por mi parte se oían palabras tipo gorda, las escuchabas? Vale, pues no era yo, y cuando veía tu cara, tu expresión pensaba up, yo no soy (It was really strange, mainly on my side because I could hear words such as fat, did you hear them? Well, it wasn't me, and when I was seeing your face, your expression, I was thinking ups, I am not)
- (242, T2) EM SAP SUPER GREU (I am really sorry)

- (242, T1) Que s'anaven colant paraules rollo idiota, fracassada, que les van gravar l'altre dia, però que no les deia pas ara eh, que les han anat colant ara però que jo no sóc així eh! (There were words heard such as idiot, loser, that were recorded the other day, but I wasn't saying them, they played them now but I am not like that ah!)
- (243, T1) jo no t'he insultat, ha sigut l'experiment, jo també estava en plan, what? (I didn't insult you, it was the experiment, I was also like, what?)
- (243, T2) Que jo no he estat eh, estava super incòmoda en plan, jod…que jo no li estic dient això, per això que HO SENTO però no he sigut jo (It wasn't me ah, I was really uncomfortable, like sh..I am not saying this, so, that's why, I am sorry, but it wasn't me)
- (248, T1) Un experimento un poco cruel, LO SIENTO (that was a cruel experiment, I am sorry)
- (248, T2) Que es un poco cruel, con los insultos y eso (It's a little bit cruel, with the insults and so)

- (250, T2) Hay palabras que te decía que yo no era, cosas que no tenían sentido, es que yo lo escuchaba y te veía y pensaba es que no lo estoy diciendo yo (There were words that I said that it was not me, things that did not make sense, I was hearing them and I was seeing you and I was thinking, I am not saying them)
- (254, T1) Me parece que has escuhcado palabras un tanto burdas y no salían de mi boca, claro no sabía si estaba permitido decirte que se escuchaban esas palabras feas, que no las decía yo (I think you heard words that were quite coarse and they did not come out from my mouth, I am not sure if I was allowed to tell you about these ugly words that were heard, that I didn't say them)

Apologies and clarifying sentences in the synchronous, neutral words condition:

- (116, T1) había unas cosas que yo no decía, no se si lo habrás notado, pero no era yo (there were some things I wasn't saying, I am not sure if you noticed, but it wasn't me)
- (116, T2) LO SIENTO… (I am sorry)
- (228, T2) Cuando decía mis palabras, que parecía que te molestaban, quizás porque estaban muy altas, no se (When I was saying the words, well it looked like they were bothering you, maybe because they were loud, I don't know)