The multimodal EchoBorg
not as smart as it looks

Falcone, Sara; Kolkmeier, Jan; Bruijnes, Merijn; Heylen, Dirk

DOI
10.1007/s12193-022-00389-z

Publication date
2022

Document Version
Final published version

Published in
Journal on Multimodal User Interfaces

Journal on Multimodal User Interfaces
https://doi.org/10.1007/s12193-022-00389-z

ORIGINAL PAPER

The multimodal EchoBorg: not as smart as it looks

Sara Falcone[1] · Jan Kolkmeier[1] · Merijn Bruijnes[2] · Dirk Heylen[1]

Received: 27 August 2021 / Accepted: 6 April 2022

Abstract
In this paper we present a Multimodal Echoborg interface to explore the effect of different embodiments of an Embodied Conversational Agent (ECA) in an interaction. We compared an interaction where the ECA was embodied as a virtual human (VH) with one where it was embodied as an Echoborg, i.e, a person whose actions are covertly controlled by a dialogue system. The Echoborg in our study not only shadowed the speech output of the dialogue system but also its non-verbal actions. The interactions were structured as a debate between three participants on an ethical dilemma. First, we collected a corpus of debate sessions with three humans debaters. This we used as baseline to design and implement our ECAs. For the experiment, we designed two debate conditions. In one the participant interacted with two ECAs both embodied by virtual humans). In the other the participant interacted with one ECA embodied by a VH and the other by an Echoborg. Our results show that a human embodiment of the ECA overall scores better on perceived social attributes of the ECA. In many other respects the Echoborg scores as poorly as the VH except *copresence*.

Keywords Embodiment · EchoBorg · Multimodality · Believability · HCI

1 Introduction

Many HCI researchers aim to create an 'artificial social entity' that is as human-like as possible, in both the (non-)verbal behaviour it exhibits and in the way its body looks. The term artificial social entities can cover a broad spectrum of technical artifacts, ranging from chatbots to virtual characters to physical social robots. In this work we focus specifically on *Embodied Conversational Agents* (ECAs). Researchers developing ECAs frequently use the Wizard of Oz (WOz, [1]) method to design and evaluate the ECA. A human operator performs the tasks of one or more components of the system that are not (yet) implemented. The person interacting with the system is tricked into believing that they are interacting with an autonomous artificial system, but in reality there is 'another person behind the curtain'. One can also imagine the complete opposite: a user is talking to a person of flesh and blood, whose decisions of what to say are made by an autonomously operating piece of software. Corti and Gillespie [2,3] introduced this WOz variant with the term *EchoBorg* (EB): a person that speaks out the utterance generated by a chatbot. This type of illusion, where a person's utterances are fully determined by a third person, was first investigated by Milgram [4] under the name *cyranic illusion*. The name refers to the French classic play *Cyrano de Bergerac* by Edmond Rostand, where the unattractive but eloquent Cyrano covertly provides the attractive Christian with the words to woo the beautiful Roxane, by whispering the right words into Christians ears from a balcony while Christian is on a date with Roxane. Milgram found that the combination of the two persons is perceived as one identity, which he named a *Cyranoid*. He investigated how different the two identities involved in the Cyranoid could be before the illusion breaks down, for instance by a child determining the utterances of an adult. Corti et al. [2] were able to maintain the cyranic illusion, even when a chatbot determines the utterances of a human the resulting EB is perceived as one identity. Confronted with a human embodiment, a user initially has no reason to question whether

✉ Sara Falcone
s.falcone@utwente.nl

Jan Kolkmeier
j.kolkmeier@utwente.nl

Merijn Bruijnes
m.bruijnes@tudelft.nl

Dirk Heylen
d.k.j.heylen@utwente.nl

[1] Human Media Interaction Department, University of Twente: Universiteit Twente, Overijssel, Netherlands

[2] Technical University of Delft, Delft, Netherlands

Published online: 05 May 2022

this person is controlled by a system. Thus, with an EB it is possible to study the 'mind' of a conversational agent without potential biases evoked by user expectations of the capabilities of an artificial agent. A user might think "it's a machine, so it won't understand me" and as such might not display, for example, conversational repair behaviour [5]. The apparatus, or *cyranic interfaces*, used by Corti and Gillespie (and before that by Milgram) are limited to the speech modality. In this paper, we present a cyranic interface for multimodal echoborgs, extending the speech-only EB method to allow for multimodal behaviour shadowing. The Multimodal EchoBorg (MEB) consist of an ECA system that dictates the speech, non-verbal back-channels, gaze and gestures of the human through a specialized interface. Using an MEB it is possible to study how all behaviours that are traditionally generated for a Virtual Human (VH) embodiment are perceived when users do not expect an artificial mind as they are interacting with a real person. We performed a study in which we compare the same interactions with an ECA that was either embodied as a VH or an MEB, both controlled by the same system. We examine the effect of the embodiment on the user perception of the agent in terms of concepts that are often used when evaluating artificial agents (i.e., animacy, anthropomorphism, intelligence) and the perception of the overall experience of the interaction.

In the next section we discuss some of the literature that looks at the perception of different embodiments. Next we describe the MEB set-up, followed by the first exploratory study.

2 Relationship between perception and embodiment

Humans interacting with others can quickly form an impression about the others' skills, personality, and attitudes towards others. These impressions can be based on just a few seconds of observing the other's appearance and (non-)verbal behavior such as facial expressions and gestures [6–9]. Effects of virtual human behaviour on perception of agent personality and interpersonal attitudes have been investigated in perceptual studies (properties of gestures [10,11] with language [12,13] on personality, posture [14] on emotion, gaze and proxemic behaviors on interpersonal attitudes [15]) as well as in studies focusing on impression shaped during first encounters with virtual characters [16].

Besides the appearance (e.g., hair colour, height), the fact that the MEB is physically embodied by a human makes it different from the VH on a screen. Li [17] discusses studies that investigate the experience of interacting with physically co-present social robots, telepresence robots and virtual agents. He concludes that *"robots were more persuasive and perceived more positively when physically present in a user's environment than when digitally-displayed on a screen either as a video feed of the same robot or as a virtual character analog"* [17, p25]. Also in human-human communication, the shape and representation of interlocutors affects how humans respond to and perceive each other. In Bailenson et al. [18], participants engaged in a technology-mediated interaction at various levels of *behavioural and form realism*, including a voice only, video conference, and through simple, virtual polygon-avatars. The reported levels of perceived co-presence and of self-disclosure were affected by those conditions. For example, both verbally and non-verbally, people disclosed more information to avatars that were low in realism. One fundamental aspect to the (M)EB is that users are (at least initially) lead to believe that they are talking with an autonomous human instead of with a machine. This is referred to as the perceived level of agency, and it is known to be an important predictor of how mediated social interactions play out. In social games, experiences are affected by beliefs about the agency of other players, and whether or not they are physically co-present. Research consistently finds that the belief that another player is human (positively) affects various aspects of the experience [19,20], such as engagement, flow, presence, enjoyment, and physiological arousal. This has also been investigated from a neuroscientific perspective: Katsyri et al. [21] found that in a first-person video game, winning versus losing activates the brain's reward circuit differently depending on the belief on whether the opponent was human or computer controlled. Concluding, a lot of evidence points towards a human, physically present interaction partner positively affects the engagement, arousal, and interactant's traits perception, over a VH on a screen.

One work that addresses the difference between how humans and agents are treated differently is that of De Melo and Gratch [22]. They propose a benchmark of believability, which according to them, requires "people, in a specific social situation, to act with the virtual agent in the same manner as they would with a real human". Based on previous research (e.g., [23,24]), they claim that the higher the attributions of mind people make, the more likely machines are to pass the benchmark of believability. Empirical evidence suggests that, compared to VHs, humans are treated more favorably in most contexts by default. The authors' theory is that this is due to the expectations we have of the other's *mind* and *experience*. Agents need to employ additional strategies and actively display capabilities to sway the user's perception of the agent in these dimensions if they seek to match a human in *believability*.

Most of the work discussed so far addresses unilateral constructs such as the flow of the experience or perceived traits of others. However, in (mediated) social interactions, there are also bilateral constructs that emerge between the interlocutors. For example, [25] have investigated *coupling* in human-agent interactions, the bilateral impact that each

interlocutor has on the other's behaviour, making the interaction a dynamic and mutual flow. According to this, both MEB and VH should exhibit the same amount of interactivity. However, we may expect that a MEB is still favored in these constructs over a VH given the overall different expectations that humans have from other humans versus from machines.

Summarizing, there is some evidence that a human (embodiment) would be favored in a number of ways over a screen-based VH embodiment - based on the physicality of the human, and based on the implied belief that a human is an autonomous conscious entity, unlike an (apparent) machine such as a VH.

2.1 How will the MEB be perceived?

Concepts and findings from the domain of mediated social interaction help us understand how the interaction with an ECA embodied by a MEB might be perceived differently from the same interaction with an ECA embodied by a VH. However, given the hybrid nature of the (M)EB (mind of a machine, body of a human), the prior work does not allow for direct predictions in this regard. In previous work on EBs, the non-EB condition featured textual interfaces rather than alternative (artificial) embodiments [2,5,26], and as such does not provide insights on how an (M)EB might perform when compared to other embodied agents. For our present work, we compare two conversational agent embodiments with a representation of a real or virtual body, pulling the compared conditions more alike. Note that our approach is not intended as the definitive study on the effect of embodiment on conversational agent perception, but intended as a first exploration of how a ECA embodied by a MEB is perceived in the dimensions relevant for our community and how sensitive the conventional measures are in this setup.

From the point of view of the methodology, we referred to Corti et. al [26] as benchmark. They analysed the adjectives participants attributed to the respective conversational partner. Participants used adjectives that are of artificial or inhuman nature ("mechanical", "computer", "robotic") to describe their interaction partner when interacting with the text interface, while used adjectives of a human nature ("shy", "awkward", "autistic") to describe the EB. Instead of asking participants to freely attribute adjectives, we administered them the commonly used Godspeed Questionnaire Series (GQS) [27] for evaluation of artificial agents. It uses semantic differential scales to cover similar concepts. These concepts are anthropomorphism, animacy, likeability, and perceived intelligence (and perceived safety, as a concept specific to robots). Given that these concepts in the GQS have a directionality from machine-like (low) to human-like (high), we expect that a human embodiment for our ECA, as achieved with the MEB, would be rated more favorably on these concepts.

Hypothesis 1 Participants will rate a MEB higher than a VH embodied conversational agent on the key concepts: anthropomorphism, animacy, likeability, and perceived intelligence.

The discussed literature demonstrates that experiences are more engaging when participants believe they are interacting with a human then when they are interacting with a machine, even if the behaviour of the other players are otherwise equal [19–21]. This depends on the bias that humans expect more relevant social actions from other humans [28]. Based on this, we would expect that the overall engagement and flow of the interaction, as well as the emotional experience and reaction, would be better experience when interacting with ECAs embodied by the MEB, rather than a VH. To rate those aspects, we administered the Game Experience Questionnaire (GEQ) [29] for the engagement and flow, and the Self-Assessment Manikin (SAM) [30], for the emotional response.

Hypothesis 2 The quality of the interaction with the ECA, as reflected in constructs such as flow, arousal and engagement (as measured by the GEQ and SAM), will be rated more positively by participants when the ECA is embodied by the MEB.

In regards to the bilateral constructs such as coupling [25], it is more difficult to make a prediction. Coupling implies an evolving equilibrium among the interlocutors. It is the capability to compensate disturbances in the interaction by evolving it. This is why it is highly complex to reproduce when employing virtual agents, since it implies that they should manage to face unexpected stimuli and situations. On the basis of the coupling concept, participants should perceive the same amount of interactivity from both human and VH embodiments. Therefore, the discourse flow and engagement should be at virtual agents level for both embodiments. However, on the basis of the reported literature, we could also assume that a MEB is favored over a VH, given the different expectations and bias that humans have from other humans and from machines, that could alter the interaction perception.

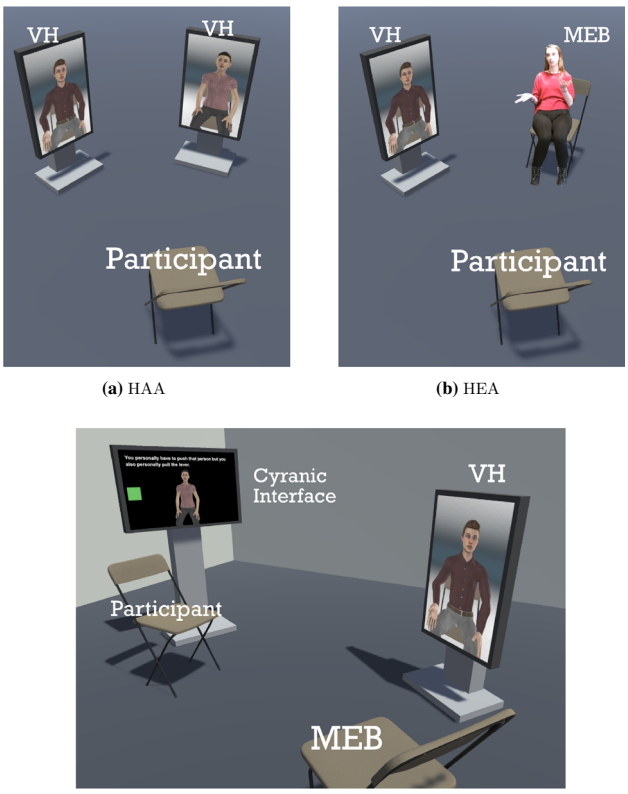
Hypothesis 3 On measures regarding the bilateral relationship between the ECA and participant during the interaction (as reflected by the coupling instrument [25]), the ECA will score higher when embodied by the MEB.

3 A cyranic interface for multimodal EchoBorgs

We designed a novel apparatus that allows for multimodal behaviour shadowing, namely speech, gestures, nonverbal back-channels, and gaze. A human shadower receives instructions of what to say and which non-verbal behaviours to display from an ECA system through the *cyranic interface* (visible in Fig. 1c).

The components of the ECA For behavior realization and planning, we employ the Articulated Social Agent Platform (ASAP) realizer [31]. For rendering the Virtual ECA embodiment on screen as well as for the cyranic display, we employ the ASAP Unity bridge [32]. The dialogue scenario is modeled using the Dialogue Game Execution Platform (DGEP) [33]. For dialogue management, we use the Flipper Dialogue Engine [34].

The Cyranic Interface The instructions to a human confederate shadowing the ECA were provided in the following way. With respect to speech, we displayed the output of our dialogue system, to be uttered by the MEB, on a screen (hidden from the participants) akin to a teleprompter. In our case, the text shown on the teleprompter was the direct output of our dialogue system, that would otherwise be spoken out by the ECA using a text-to-speech (TTS) system. We explored the alternative to play audio of the utterances through hidden earpieces, more similar to the conventional speech shadowing. However, it appeared to be very difficult to shadow a TTS voice. Moreover, while the utterance selection of the system is dynamic, the ECA utterances in our user study were pre-scripted. After a bit of practice, our MEB became familiar with the utterances, and managed to shadow the speech fluently.



(a) HAA **(b)** HEA

(c) View of the echborg in HEA

Fig. 1 3D illustration of the debater placement in **a** HAA and **b** HEA conditions. The echoborg view in **c** shows how the screen with the cyranic interface was positioned, behind the participant

A simple ECA gaze behaviour model sufficed as we envisioned a triadic interaction. Therefore, we could keep the interface for gaze shadowing simple: there is a green highlight at the left or right half of the screen, indicating whether gaze should be directed to the conversation partner on the left or right (from the perspective of the MEB).

The Echoborg was also instructed to back-channel at certain times While listening. Our ECA system only includes a single type of back-channel, head nods. In the MEB interface, these behaviors are signaled by (discretely) flashing the word *nod* on the screen.

When it comes to gestures, shadowing motion and poses are challenging for the MEB. Lexical instructions for gestures are difficult to translate into fluent and animate motions that retain the semantic connection with the words uttered. As an alternative, we decided to show the motions on an animated copy of the ECA, rendered on the screen behind the participant. While ad-hoc mimicking remained difficult, we observed a learning effect, as with the speech shadowing. Because the speech and gestures generated by the system were the same for each utterance, our MEB was able to learn the speech and gestures produced by our system and was able to shadow with similar 'size' and 'stroke' from seeing the animation only in peripheral vision.

4 Exploratory user study

Unlike the prior work on EBs with unscripted dialogues [2], we modeled more strictly the dialogue scenario for our ECA. Besides the increased experimental control when comparing the interaction between embodiments, it also simplifies the complexity of the overall system.

We modelled an ethical-debate-like scenario, with a *moderator* and two opposing debaters, the *proponent* and *opponent* discussing different variations of the *Trolley Dilemma* [35]. It is an ethical dilemma questioning about whether to sacrifice one person to save a larger number. The scenario is a runaway trolley barreling down the railway tracks. On the tracks there are five people tied up and unable to move and the trolley is headed straight for them. A person is standing in the train, next to a lever. Pulling the lever, the trolley will

switch to a different set of tracks. However, there is one person on the side track. The proponent is asked to argue for pulling the lever, while the opponent is asked to argue for staying passive. The moderator's role is to open and manage the discussion and to introduce the dilemma and its variants to the debaters before yielding the floor to them for their arguments.

4.1 Modelling ECA and dialogue

To model this scenario and to inform the design of utterances and gestures for the ECAs, all roles (debaters and moderator) are modelled from an only-humans debate corpus. We recorded audio and video, and we transcribed the dialogues. We also measured some of the interaction experience and interlocutor perception that were also used in the user study later on.

In total, we recorded 6 triads (2 females, 16 males). From the transcriptions, the arguments used to defend the two debaters' positions (pulling the lever/being passive) were categorized by type of argument (see Table 1) and selected for our ECA to use as utterances. In a small survey (20 participants on SurveyMonkey), external raters ranked the selected utterances for the different arguments based on their strength to convince. This allowed us to select balanced arguments for both proponent and opponent. For the non-verbal behaviors of the ECA, we have consulted the video recordings from the corpus of the selected utterances and presented them to an actor. The actor acted out these utterances wearing a MoCap suit. This yielded full-body gesture animations for each utterance. The MoCap recordings and selected utterances were then combined and linked to the dialogue move. As mentioned in Sect. 3, our ECA system uses the DGEP dialogue argumentation framework. DGEP uses the concept of *dialogue moves*, namely the schematic representations of a single move in a dialogue, its reply and the connections to the argument structure.

Table 1 The key arguments, that we classified, of the *Trolley Dilemma* debate and the moral questions which describe them

Key arguments	Moral question
Fate	Can the fate decide for the life of human beings?
Numeric	Human life is a qualitative or quantitative matter?
Economic	Is it better to save more life because they are a greater resource for the society?
Responsibility	If we make the choice of pulling the lever, do we become responsible of a murder?
Inaction	Can 'inaction' be considered as 'action'?

4.2 Experiment design

Participants were assigned to one of two conditions: *Human-Agent-Agent* (HAA) or *Human-Echoborg-Agent* (HEA). Participants were always assigned to the role of the moderator, while the debaters (*proponent* and *opponent*) were always acted out by our ECAs. In both HAA and HEA, the opponent was always embodied by the VH. In HEA, the proponent was embodied by the MEB, while in HAA, the proponent was also embodied by a VH. We call this between-subject variable *proponent embodiment*. For those participants assigned to the HEA condition it is also interesting to compare their ratings of the VH embodiment of the opponent versus the MEB proponent embodiment. This is a within-subject variable which we refer to as *debater embodiment*.

4.3 Materials and apparatus

The moderator and the two debaters are positioned in a triangle (see Fig. 1a and b). VHs were shown on large TV screens in portrait mode. When the proponent was embodied by the MEB, that screen was replaced by a chair for the MEB to sit on. For the MEB's cyranic interface, a large screen was placed behind and out of sight off the participant, facing the MEB (see Fig. 1c). Due to the fact that there were other screens in the experiment room, participants did not get suspicious in seeing the screen behind their chair while entering in the room. Moreover, all the screens were, or appeared as, turned off when participants entered the room. Therefore, they could not see the agent on the screen.

The moderators received cue-cards to guide the debate through the different variants (in any order). The cue-cards represented utterance hints that participants could rephrase and use in the order that they preferred while interacting with the two debaters. This allows for the participant to partake in the interaction without affecting the conversation in an unpredictable way.

The detection of when the participant is speaking, and which move their utterances represent, is done secretly by the experimenter in a WOz fashion [1].

4.4 Multimodal EchoBorg training

We recruited an experienced actress from the student body to act as the MEB in this user study (see Fig. 1b). Following a number of training sessions of the debate with the researchers, she became familiar with the scenario and behaviours. While not systematically quantifying the accuracy of shadowing, comparing recordings of MEB behaviors with the VH behaviours showed that the actress was able to shadow the speech and gestures reliably.

4.5 Participants

The Ethics board of the [Anonymous] approved the user study. In total, 36 participants were sampled from the university staff and student body, between 19 and 46 years old, 16 females and 20 males, and the number of participants was equally distributed between conditions.

4.6 Measures

We selected several existing questionnaires measuring interaction experience and interlocutor's perception that are commonly used in the IVA community, as discussed in Sect. 2.1. Therefore, we used the GQS to address the first hypothesis, which concerns the effect of the appearance, and the virtual or physical presence of the embodiment on the human interlocutor's perception. To address the second hypothesis, related to the effect of the embodiment on the interaction experience, we had participants fill out the Game Experience Questionnaire (GEQ) [29] and the Self-Assessment Manikin (SAM) [30]. Finally, to address the third hypothesis, we measured the dynamic *coupling* between the participants and the ECA embodied by both the VH and the MEB using the questionnaire from [25].

5 Results

We conducted a one-way ANOVA on the effect of the between subject variable "proponent embodiment" for each of the sub-scales of the questionnaires and dimensions described above. Two of the GQS sub-scales showed significant effects: animacy ($F(1,34) = 5.834$, $p = 0.021$, $\eta^2 p = 0.146$) and anthropomorphism ($F(1,34) = 20.061$, $p < 0.001$, $\eta^2 p = 0.371$). Post-hoc tests show that the proponent was rated higher on animacy and anthropomorphism, when embodied by the MEB. On the co-presence sub-scale of the coupling questionnaire, we found a significant effect of the "proponent embodiment" between configurations ($F(1,34) = 16.920$, $p < 0.001$, $\eta^2 p = 0.332$). Post-hoc tests revealed that the proponent was rated higher on co-presence, when embodied by the MEB. Since participants within the Human-EchoBorg-Agent (HEA) condition ($n = 18$) interacted with both an MEB and a VH embodiment, we conducted an ANOVA on the effects of the within subject variable "debater embodiment" on those sub-scales that measure attributes of the individual debaters. Again, two sub-scales showed (near) significant effects: anthropomorphism ($F(1,17) = 12.190$, $p = 0.003$, $\eta^2 p = 0.418$) and perceived intelligence ($F(1,17) = 4.322$, $p = 0.053$, $\eta^2 p = 0.203$). There were no other significant effects of "proponent" and "debater embodiment" found on any other sub-scales. Statistics for the between and within post-hoc tests are reported in Table 2, and response distributions are visualized in Figs. 2, 3 and 4.

6 Discussion

Reiterating, we compared participants' perception of a traditional VH embodiment with a MEB embodiment, while both had the same conversational agent ('mind') determining the utterances and behaviour they display during a debate. We examined the participants' perception of the agent in terms of concepts that are often used when evaluating artificial agents, and participants' perception of the overall experience of the interaction.

6.1 Comparing the multimodal EchoBorg and virtual human embodiments

Looking at the hypotheses, we observe the following. We only partially support our first hypothesis, that the MEB is perceived as more favorably than the VH on perceived agent traits: while the MEB does score higher on the Godspeed instrument sub-scales that measure the anthropomorphism (both in the within and between comparison) and animacy (between subjects), there is only near-significance for the intelligence in the within comparison, and no difference in likability ratings. These results suggest that interaction is more than just appearance. Our interpretation here is that only measures that relate to the outer appearance of the embodiment seem to be favored by the human embodiment, while it fails to lead participants into (falsely) overestimating traits that are related more to the behaviour of the conversation partner - i.e. the intelligence and likability.

Our second hypothesis, the quality of the interaction with an MEB will be rated more positively than with a VH, is rejected. We had speculated that whenever there is another human involved, even though it displays the same limited behaviours and interactivity as displayed on the virtual embodiment, the interaction would be perceived as more engaging and interesting. This appears not to be the case, as interactions featuring the MEB were not rated more positive than those only featuring VH embodiments. Together with the observation in regards to the first hypothesis, this may lead us to assume that any initial expectation favoring a human embodiment are overruled by the limited perceived mind during the interaction.

Finally, our third hypothesis concerns the how participants perceived their bilateral relationship with the ECA. We hypothesized that the MEB would be rated more favorably, because the human appearance evokes the expectation of a human level of interactivity. Based on our results, we reject this hypothesis. Looking in more detail at the sub-scales, coupling with the debater, engagement and believability did not

Table 2 Statistics of pairwise comparisons

	Scale	Subscale	Contrast	Estimate	SE	df	*t*.ratio	*p*. value
Between	Godspeed	Animacy	VH–EB	−0.722	0.299	34	−2.415	0.021
	Godspeed	Anthropomorphism	VH–MEB	−1.233	0.275	34	−4.479	0.000
	Coupling	Copresence	VH–MEB	−0.639	0.155	34	−4.113	0.000
Within	Godspeed	anthropomorphism	VH–MEB	−1.022	0.293	17	−3.491	0.003
	Godspeed	Intelligence	VH–MEB	−0.537	0.258	17	−2.079	0.053
	Coupling	Copresence	VH–MEB	−0.667	0.133	17	−5.030	0.000

Bottom half showing the comparisons of scores attributed to the proponent (VH) and the opponent (MEB) within the HEA condition.Top half showing the comparisons of scores attributed to the proponent debater embodiment (VH or MEB) between subject

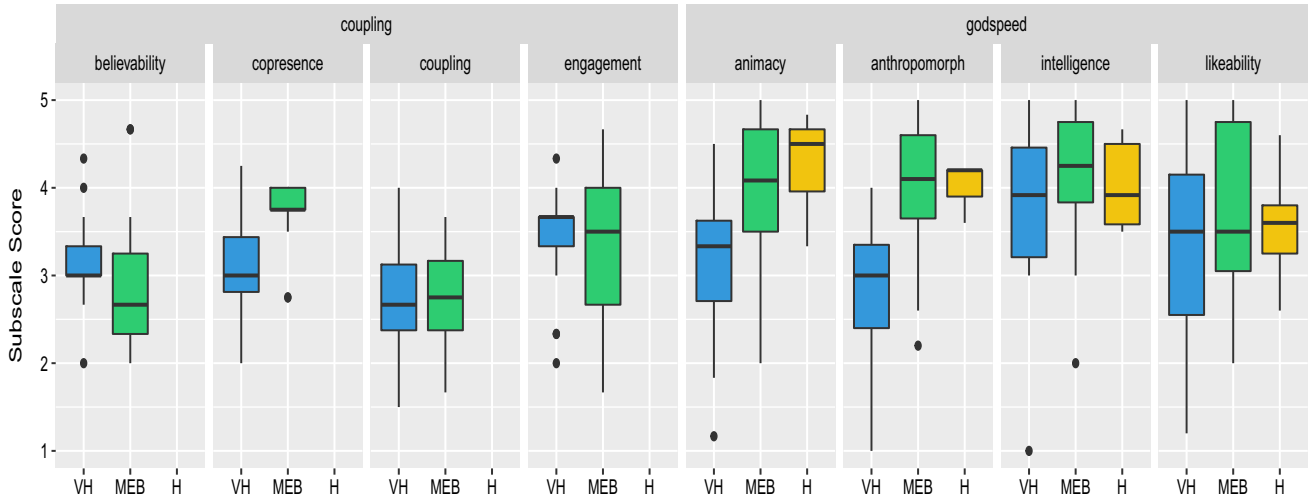


Fig. 2 Moderator scores attributed to the proponent debater embodiments (between subject), also showing the moderator scores for the proponent in the Human-only pre-study corpus

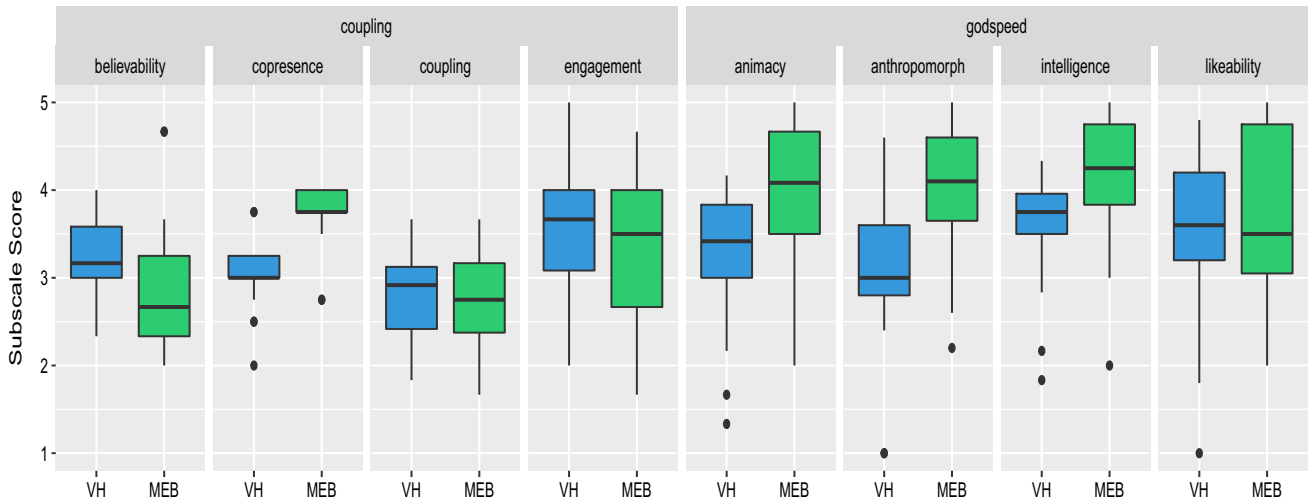
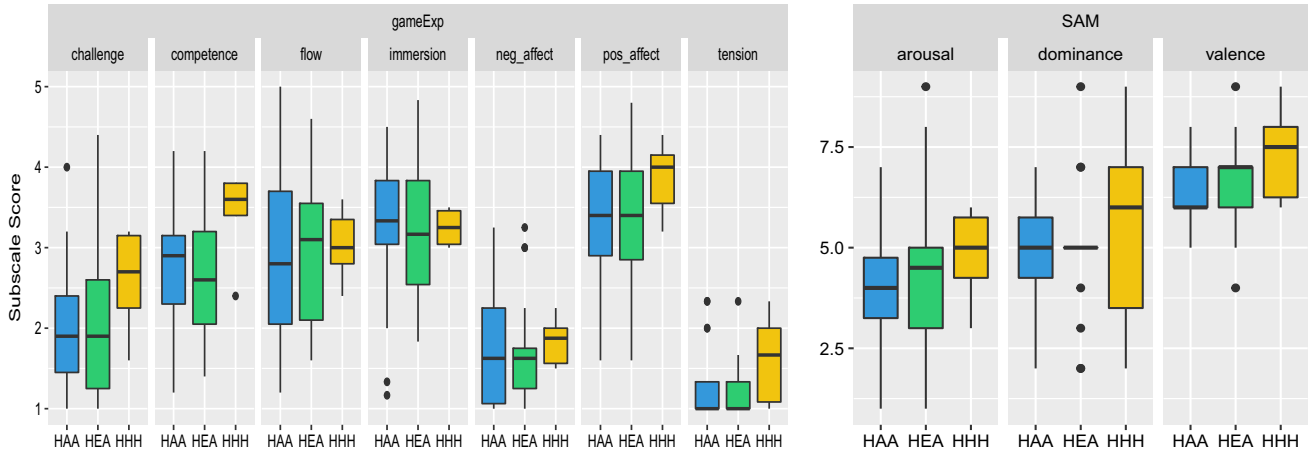


Fig. 3 Moderator scores attributed to the different debaters (within subject) in the HEA setting (virtual human acting as opponent, EchoBorg acting as proponent)

score significantly higher for the MEB. Only the co-presence sub-scale the MEB was rated significantly higher. This is a measure that might be more influenced by the physicality of the embodiment rather than by the displayed behaviour. Thus, a human embodiment might not create a better rela-



(a) Game Experience Questionnaire **(b)** SAM

Fig. 4 Scores on moderator game experience (**a**) and SAM self-report scales (**b**) between the different combinations of debaters (HAA = only agents, HEA = Multimodal EchoBorg as proponent, HHH = human only pre-study corpus)

tionship between a user and ECA, but might evoke higher feelings of co-presence.

In an attempt to find alternative explanations, we may consider works such as that of Nowak and Biocca, who found that more anthropomorphic embodiments of agents (or in their case avatars) might "set up higher expectations that lead to reduced [co-]presence when these expectations were not met" [36, p481]. Initially, a MEB will set up the highest expectations, while the limited capabilities of our conversational agent very likely meant that the MEB was not be able to meet those expectations during the interaction.

It is also important to consider one other limitation of this study, namely the sample size. Its small dimension could under power the statistical significance of the results. We need to replicate the experiment with a larger population. However, the present study still shows a possible methodology and how sensitive the conventional measures are in a setup like this. The reported results are not the main contributions, but they provide an overview on the effects that the MEB can have on a human interactant in this preliminary version.

6.2 Comparisons with a human-only experience

Next we explore the scores of the different ECA embodiments with the scores collected in the all-human corpus recording sessions. We find that the scale ratings attributed to human proponents, for the GQS, are quite similar to those attributed to the MEB proponents on all sub-scales (see Figure 2). While we expected this for the perceived animacy and anthropomorphism sub-scales, we also expected the humans to receive much higher ratings on intelligence and likeability, based on the coupling concept [25]. The experience in the pre-study corpus, in fact, was more open and interactive than the experiment sessions. Due to the fact that all the interactants were participants, and there was not a virtual agent limiting the conversation or creating bias. Instead, we find that the levels are similar to both the VH and the MEB ratings. For intelligence, an explanation may be a ceiling effect, with medians and upper quartiles concentrating around the 4–5 point level of the sub-scale. For the GEQ, comparing the responses of moderators from the human-only pre-study corpus to the responses in the experiment sessions, we see a different trend from the debater perception rating discussed before (see Fig. 4a). The experience from the human-only session scores seem much higher in terms of perceived challenge, competence, positive affect and tension when compared to the experiment sessions. Similarly on the SAM-instrument, the ratings on arousal and valence seem somewhat higher (on dominance we have a high variance in the responses, but the median level is also higher). Thus, perhaps the increased interactivity of the human debaters informed these measures—which would further support that the limiting factor for the MEB scores are based on the limitations of the ECA system controlling the MEB, which the human embodiment could not hide. Alternatively, the human corpus recording sessions had different rules and featured a less structured debate, which may also have affected the game experience scores. During those debate sessions, social dynamics and unexpected stimuli were more common. On the basis of the literature, this probably contributed to increase the level of attention, arousal, and engagement.

6.3 An evaluation and inner perspective of the multimodal EchoBorg

A contribution of this work is the first implementation of a Multimodal EchoBorg apparatus for ECA systems. To understand the limitations and how to improve it in the future, we asked the participants, at the end of the experiment,

to provide a feedback on the MEB interlocutor before we revealed that the actress who acted as the MEB was a confederate. All the participants reported that, after more or less five minutes of conversation, they perceived the interlocutor as awkward. They provided different explanations for this behaviour: some participants thought that the interlocutor was shy, others thought that the interlocutor had some mild form of mental disorders, only two participants understood that she was a confederate and she was acting. We also asked the actress that acted as the MEB to fill out a self-report after each session. She reported deviations from the behaviour that the ECA asked her to perform. Specifically, she reported how, when and why she deviated and in which modality she deviated (listening behaviour, speech, gestures). We compared her reported deviations with recordings of her actual behaviour to check if her perception was consistent to the real experience. The actress never reported deviations in the listening behaviour. She reported most deviations for speech, citing as reason:

(i) *"I thought that was the sentence I had to say but instead I said it faster."*; (ii) *"I tried not to look at the screen because I felt that the participant might notice something is happening behind him."*; (iii) *"The participant wanted to speak and I had to speak over him."*. Concerning the gestures, the actress reported that it was not always easy to shadow the gestures from the interface, for example: *"I had the impulse to follow my own reaction to what I was saying"*. From the recordings, we observed that the majority of deviations happened during the gesture shadowing, while we observed only very small variations in the speech shadowing, and no variations in the listening behaviours. Thus, the actresses self-reported deviations and the observed deviations were not in line, suggesting that the actress was perhaps less aware of her performance on gesture shadowing. Perhaps integrating an automatic feedback mechanism of shadowing behavior in a future MEB setup could improve the quality of shadowing.

7 Conclusion and future works

We explored how the embodiment of an ECA influences the perception of the interaction using an upgraded version of the EchoBorg method, the Multimodal EchoBorg. We present our first experiences of employing the EB method in ECA research. From a practical standpoint, we have built an apparatus for multimodal shadowing, and gained insights in how it can be employed with a confederate in an experimental setting. From the user-study, we have obtained a first overview on the biases that may occur when replacing the embodiment of a VH with a real human, keeping all other aspects of the ECA behavior the same. In summary, the results from our study do not support our initial assumption that an experience with an MEB would always be rated favorably over the same interaction with a VH based on the belief that (one of) the actor(s) was a human. Instead, we see that the limited artificial mind may shine through more than expected, limiting such favorable ratings.

We acknowledge a number of limitations of the present work. First and foremost, the sample size was relatively small for ANOVA with post-hoc tests. We reported significant results, however the study also has a possibly inflated test power due to the procedure used. Moreover, the study design lacks counterbalancing in debater role and gender, and the analysis of both within and between subject comparisons in this way may have inflated statistical power. Future studies are necessary and may benefit from a different study design. For example, a dyadic interaction scenario with a strict between subject design is more suitable for a more rigorous investigation of the MEB when studying perception biases. Furthermore, metrics for the shadowing performance of the MEB need to be defined and measured for control purposes. The next important step to understanding if and how we can benefit from the MEB method for ECA development is to look more at how the user is treating the MEB, perhaps with a similar methodology as the one used in [5]. Additionally, there are possibilities to improve the MEB interface further, allowing for more accurate shadowing in even more modalities using, for example, visual overlays in covert AR glasses, or perhaps haptic displays that provide information for motion in different bodyparts.

In fact, we recognize that in our study, the MEB was potentially over-reliant on apriori knowledge of the dialogue. She was able to practice her performance, as in large parts speech and the accompanying gestures were fully deterministic. For a future *production* MEB system, also dynamic, spontaneous behaviours should be possible to realize. Additionally, not all MEB behaviours could be controlled (e.g., nonverbal leakage). There may even be systematic biases that are not controlled for, for example in the MEB's gaze behavior, due to the use of the MEB interface.

Reflecting on Rostand's play *Cyrano de Bergerac*, the moral of the story was that Roxane was attracted to Christian's body, but ultimately fell in love with Cyrano's mind: a feat not likely repeated by our MEB, as our ECA *indeed* turned out to be not as smart as it looked.

Funding There is no founding to report for this submission.

References

1. Dahlbäck N, Jönsson A, Ahrenberg L (1993) Wizard of Oz studies—why and how. Knowledge-Based Syst 6(4):258–266. https://doi.org/10.1016/0950-7051(93)90017-N
2. Corti K, Gillespie A (2015) A truly human interface: interacting face-to-face with someone whose words are determined by a computer program. Front Psychol 6:634. https://doi.org/10.3389/fpsyg.2015.00634
3. Gillespie A, Corti K (2016) The body that speaks: recombining bodies and speech sources in unscripted face-to-face communication. Front Psychol 7:1300
4. Milgram S, van Gasteren L (1974) Das Milgram-Experiment.) Rowohlt
5. Corti K, Gillespie A (2016) Co-constructing intersubjectivity with artificial conversational agents: people are more likely to initiate repairs of misunderstandings with agents represented as human. Comput Human Behav 58:431–442
6. Ambady N, Rosenthal R (1993) Half a minute: predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. J Personal Soc Psychol 64(3):431
7. Campbell A, Rushton JP (1978) Bodily communication and personality. Br J Soc Clin Psychol 17(1):31–36
8. Levesque MJ, Kenny DA (1993) Accuracy of behavioral predictions at zero acquaintance: a social relations analysis. J Personal Soc Psychol 65(6):1178
9. Argyle M (2013) Bodily communication. Routledge
10. Neff M, Toothman N, Bowmani R, Tree JEF, Walker MA (2011) Don't scratch! self-adaptors reflect emotional stability. In: International workshop on intelligent virtual agents, Springer, pp 398–411
11. Smith HJ, Neff M (2017) Understanding the impact of animated gesture performance on personality perceptions. ACM Trans Graph (TOG) 36(4):49
12. Neff M, Wang Y, Abbott R, Walker M (2010) Evaluating the effect of gesture and language on personality perception in conversational agents. In: International conference on intelligent virtual Agents, Springer, pp 222–235
13. Di Maro M, Falcone S, Cutugno F (2018) Prosodic analysis in human-machine interaction. Studi AISV 1
14. Normoyle A, Liu F, Kapadia M, Badler NI, Jörg S (2013) The effect of posture and dynamics on the perception of emotion. In: Proceedings of the ACM symposium on applied perception, ACM, pp 91–98
15. Kolkmeier J, Vroon J, Heylen D (2016) Interacting with virtual agents in shared space: Single and joint effects of gaze and proxemics. In: International conference on intelligent virtual agents, Springer, pp 1–14
16. Cafaro A, Vilhjálmsson HH, Bickmore T, Heylen D, Jóhannsdóttir KR, Valgardsson GS (2012) First impressions: users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In: International conference on intelligent virtual agents, Springer, pp 67–80
17. Li J (2015) The benefit of being physically present: a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. Int J Human-Comput Stud 77:23–37
18. Bailenson JN, Yee N, Merget D, Schroeder R (2006) The effect of behavioral realism and form realism of real-time avatar faces on verbal disclosure, nonverbal disclosure, emotion recognition, and copresence in dyadic interaction. Presence: Teleop Virtual Environ 15(4):359–372
19. Weibel D, Wissmath B, Habegger S, Steiner Y, Groner R (2008) Playing online games against computer-vs. human-controlled opponents: Effects on presence, flow, and enjoyment. Comput Human Behav 24(5):2274–2291
20. Lim S, Reeves B (2010) Computer agents versus avatars: responses to interactive game characters controlled by a computer or other player. Int J Human-Comput Stud 68(1–2):57–68
21. Kätsyri J, Hari R, Ravaja N, Nummenmaa L (2013) The opponent matters: elevated fmri reward responses to winning against a human versus a computer opponent during interactive video game playing. Cerebral Cortex 23(12):2829–2839
22. de Melo CM, Gratch J (2015) Beyond believability: Quantifying the differences between real and virtual humans. In: Brinkman W-P, Broekens J, Heylen D (eds) Intelligent virtual agents. Springer, Cham, pp 109–118
23. Blascovich J, Loomis J, Beall AC, Swinth KR, Hoyt CL, Bailenson JN (2002) Immersive virtual environment technology as a methodological tool for social psychology. Psychol Inq 13(2):103–124
24. Blascovich J, McCall C (2013) Social influence in virtual environments
25. Bevacqua E, Stanković I, Maatallaoui A, Nédélec A, De Loor P (2014) Effects of coupling in human-virtual agent body interaction. In: International conference on intelligent virtual agents, pp 54–63. Springer
26. Corti K, Gillespie A (2015) Offscreen and in the chair next to your: conversational agents speaking through actual human bodies. In: Intelligent virtual agents, pp 405–417 . Springer
27. Bartneck C, Kulić D, Croft E, Zoghbi S (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Robot 1(1):71–81
28. Okita SY, Bailenso J, Schwartz DL (2007) The mere belief of social interaction improves learning. In: Proceedings of the annual meeting of the cognitive science society, Vol 29
29. IJsselsteijn W, De Kort Y, Poels K (2013) The game experience questionnaire
30. Bradley MM, Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. J Behav Therapy Exper Psychiatry 25(1):49–59
31. Van Welbergen H, Reidsma D, Kopp S (2012) An incremental multimodal realizer for behavior co-articulation and coordination. In: International conference on intelligent virtual agents, pp 175–188 . Springer
32. Kolkmeier J, Bruijnes M, Reidsma D, Heylen D (2017) An asap realizer-unity3d bridge for virtual and mixed reality applications. In: International conference on intelligent virtual agents, pp 227–230. Springer
33. Lawrence J, Snaith M, Konat B, Budzynska K, Reed C (2017) Debating technology for dialogical argument: sensemaking, engagement, and analytics. ACM Trans Internet Technol (TOIT) 17(3):1–23
34. van Waterschoot J, Bruijnes M, Flokstra J, Reidsma D, Davison D, Theune M, Heylen D (2018) Flipper 2.0: a pragmatic dialogue engine for embodied conversational agents. In: Proceedings of the 18th international conference on intelligent virtual agents. IVA '18, pp. 43–50. ACM, Sydney, NSW, Australia. https://doi.org/10.1145/3267851.3267882
35. Thomson JJ (1976) Killing, letting die, and the trolley problem. The Monist 59(2):204–217
36. Nowak KL, Biocca F (2003) The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. Presence: Teleop Virtual Environ 12(5):481–494