future internet

MDPI

Article

Enhancing Human–Agent Interaction via Artificial Agents That Speculate About the Future

Casey C. Bennett [1,2,*], Young-Ho Bae [2], Jun-Hyung Yoon [2], Say Young Kim [3] and Benjamin Weiss [4]

[1] School of Computing, DePaul University, Chicago, IL 60604, USA
[2] Department of Data Science, Hanyang University, Seoul 04763, Republic of Korea; byh711@hanyang.ac.kr (Y.-H.B.); junehyung96@gmail.com (J.-H.Y.)
[3] Department of English Language & Literature, Hanyang University, Seoul 04763, Republic of Korea; sayyoungkim@hanyang.ac.kr
[4] Quality and Usability Lab, Technische Universität Berlin, 10623 Berlin, Germany; benjamin.weiss@tu-berlin.de
* Correspondence: cbenne33@depaul.edu

Abstract: Human communication in daily life entails not only talking about what we are currently doing or will do, but also *speculating* about future possibilities that may (or may not) occur, i.e., "anticipatory speech". Such conversations are central to social cooperation and social cohesion in humans. This suggests that such capabilities may also be critical for developing improved speech systems for artificial agents, e.g., human–agent interaction (HAI) and human–robot interaction (HRI). However, to do so successfully, it is imperative that we understand how anticipatory speech may affect the behavior of human users and, subsequently, the behavior of the agent/robot. Moreover, it is possible that such effects may vary across cultures and languages. To that end, we conducted an experiment where a human and autonomous 3D virtual avatar interacted in a cooperative gameplay environment. The experiment included 40 participants, comparing different languages (20 English, 20 Korean), where the artificial agent had anticipatory speech either enabled or disabled. The results showed that anticipatory speech significantly altered the speech patterns and turn-taking behavior of both the human and the agent, but those effects varied *depending* on the language spoken. We discuss how the use of such novel communication forms holds potential for enhancing HAI/HRI, as well as the development of mixed reality and virtual reality interactive systems for human users.

Keywords: human–robot interaction; social cognition; virtual avatar; speech system; language differences; virtual reality

check for updates

Academic Editors: Rui Yu, Sooyeon Lee, Syed Masum Billah and John M. Carroll

Received: 13 December 2024
Revised: 8 January 2025
Accepted: 18 January 2025
Published: 21 January 2025

Citation: Bennett, C.C.; Bae, Y.-H.; Yoon, J.-H.; Kim, S.-Y.; Weiss, B. Enhancing Human–Agent Interaction via Artificial Agents That Speculate About the Future. *Future Internet* **2025**, *17*, 52. https://doi.org/10.3390/fi17020052

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Overview

It could be argued that most of what humans talk about is not actions we necessarily do ourselves, but rather *speculation* about events that happen in the world around us. The weather, the economy, politics, sports, etc. This includes events that have occurred in the past as well as events that may (or may not) occur in the future, i.e., things we imagine as rational possibilities [1]. Natural human conversation is full of this kind of speculation, which may eventually lead to actions but is not yet action in and of itself. In other words, it is what we might refer to as "anticipatory speech", a term we coin here to define it. In contrast to the more typical planning approach for artificial agent speech, which is geared towards specific actions the agent will perform, the focus of anticipatory speech is on potential events with unclear probability. Humans are thought to engage in such speech because it

allows them to cognitively structure their own thoughts by mentally rehearsing various actions toward multiple potential events [2,3], while also communicating a sense of "shared experience" to other people around them in order to create the social cohesion necessary for surviving in an uncertain world (i.e., "we're all in this together" mentality) [4–7]. The purpose of anticipatory speech in that sense is very different from the purpose of speech planning, in that for the former, we are not really planning actions but rather attempting to engage socially. As such, we would argue that perhaps our social robots and artificial agents should do the same sort of thing, given that they were inherently designed to be social (i.e., unlike service robots or physical machines). Intuitively, that makes sense, but the challenge, of course, is how to accomplish it in practice [8,9]. After all, it is easy enough to talk about what is happening, or has already happened, or even what we will definitely do in the future, but in contrast, talking about what *might* happen and what we *might* do is much more abstract.

It is important to clarify the difference between the definition of anticipatory speech given above and other related concepts: joint planning, speech planning, dialogue state trackers, turn-taking prediction, etc. Joint planning, for instance, focuses on the actions that an agent/robot and human will perform towards some specific task (i.e., no speculation involved) in an attempt to create better coordination between them. It can be thought of as an extension of classical planning methods in computer science and AI [10–13]. Speech planning is a term originally from neuroscience that refers to the human production of speech content (i.e., taking an idea and turning it into words) and is often applied to populations where such speech production is disrupted (e.g., people who stutter, post-stroke patients) [14–16]. Dialogue state trackers are used in many modern speech systems, but typically focus on what has already been said in order to try to keep the agent/robot "on-topic" during the conversation [17–19]. Turn-taking prediction focuses on determining appropriate times for agent/robot to speak so as not to interrupt a human conversational partner [20,21]. As one can see from such descriptions, all those other terms have a different meaning and purpose from anticipatory speech, which is focused on *speculating* about potential events.

Presently, many speech systems for human–agent interaction (HAI) and human–robot interaction (HRI) focus on current actions and/or knowledge of prior events (see Section 2), and even large-language model (LLM) systems like ChatGPT can be framed similarly given their corpus-based training methods [22,23]. To our knowledge, there is limited research on embedding speculation into speech systems for artificial agents. A search on Google Scholar for "speculation" along with terms such as "conversational agent", "conversational user interface", "speech system", etc., returns only a handful of relevant articles, e.g., [24–26], though to be clear, this paper is not intended as a full-blown meta-review of the literature. Moreover, many of those papers did not involve actual experiments with a robot/virtual agent, but rather were theoretical papers (in contrast to this paper). As such, a good first step towards this idea is to explore different methods for creating "anticipatory speech" in artificial agents and measure the effects thereof. Towards that end, we designed a human interaction experiment with an autonomous conversational agent (in the form of a 3D virtual avatar) that could interact with humans in a cooperative survival game environment. The avatar was equipped with anticipatory speech capabilities, which could be turned on and off to test the effects. Our primary aim was to see if there were any differences during interaction with an artificial agent that had anticipatory speech enabled versus one that had it disabled. Additionally, we wanted to see if any detected differences were consistent across different languages or varied across languages.

This research has potential applicability to HAI, HRI, and more broadly, conversational user interfaces (CUIs) in terms of augmented and mixed-reality systems that attempt to

integrate virtual interaction and physical interaction within a single platform [27,28]. We discuss this potential in the Section 5 at the end.

2. Background

2.1. Theoretical Work

The question of what makes a conversational agent or interface "intelligent" is a complicated one [29]. From a theoretical perspective, the broad notion that such intelligence would entail having an artificial agent that could anticipate events in the real world before they occur has a long history [30,31]. More recently, there have been theoretical attempts to address it in the fields of HAI and HRI, some of which have applicability to the problem we are addressing here.

For instance, Chater (2023) argued that the concept of social cohesion (introduced in Section 1) should be seen more as "virtual bargaining" between two parties. In short, the collective meaning inherent in social communication is not defined by any individual person (or artificial agent for that matter), and as such, communicating about the future should be viewed as a sort of "negotiation" between conversational partners about what is important to focus on (versus what is not), based on known social norms [32]. After all, humans are not really free to do "whatever they want", even though there is a lot of flexibility within the constraints of social and cultural systems. Sometimes we do not even clearly articulate our speech on purpose (e.g., dialect use, slang), in order convey hidden meaning or signal status to others [7]. As such, we may be placing too heavy a burden on our artificial agents by omitting the fact that what we say is not always the "real point" when we engage in efforts to create more natural AI interactions by focusing on what our artificial agents say, since there is so much being sub-communicated in real human interactions that has to be interpreted by both parties. In other words, it is not a fixed set of "rules" that govern human behavior, but rather a *negotiation* about the future that determines the "rules" in the real world [32]. Speculation, in that sense, is a form of negotiation. Thus, if an artificial conversational agent can negotiate the rules through speculation, then it may possibly have a greater chance of successful interaction with humans. Suffice it to say, there are some intriguing parallels between our notion of speculating about the future via anticipatory speech and Chater's concept of social negotiation.

Beyond the above, some researchers have recently looked into creating virtual simulations of cooperation between human infants and their caregivers (including both verbal and non-verbal interaction) as a way of understanding simple cooperative scenarios prior to the social complexity we typically see among human adults [33,34]. The idea is that developing a stronger theoretical basis through such virtual simulation will identify the basic components of social cooperation that could then potentially serve as building blocks for improved artificial agents and robots. The strategy of using these kinds of virtual simulations to model more complex social scenarios has some similarities to this research here, though our work is more experimental in nature (see the next section).

Of course, there are other studies in existence related to coordinated joint actions between humans and robots conducted in lab settings meant to understand the theoretical basis of physical coordination, but those focus mainly on the mechanics of the physical action itself (e.g., kinematics) [35]. Though such coordination can at times involve speech communication, that is not the focus, nor are those studies designed specifically to study speech interaction [10,36]. As such, that research has limited applicability to our specific problem scenario here, which does not necessarily require any sort of physical interaction to occur. Indeed, user models in HRI/HAI cover many modes of interaction with social robots and virtual avatars, not all of which are physical in nature, even when applied to a physically embodied robot [37].

2.2. Experimental Work

Aside from theoretical work, there has also been prior experimental research on speech systems for artificial agents geared towards future actions, which, although different from the concept of anticipatory speech here, still provides some useful background on how scientists are exploring different types of speech with regard to artificial agents.

For instance, in the field of HRI, many researchers have evaluated the effects of speech on trustworthiness and user perception, though not necessarily the effects on actual task performance [13,38]. It remains an open question whether impacting trust will actually carry through to impact completing some task jointly by a human and artificial agent in the real world, though that is currently a hot area of research [39]. There have also been a few attempts to measure the effects of verbal feedback from a robot/agent prior to human-performed actions during interactions, though the results of such research have been mixed. Some researchers report significant effects, while others report no clear effect; e.g., the human-performed action remains the same after feedback [40–43]. That may be an issue with how the feedback is being communicated by the robot that limits its impact or have something to do with user perceptions of the robot's intelligence [44]. Suffice it to say, while there is a general consensus in the field that trust and alignment are important factors in HRI, it remains unclear as to what degree.

Other researchers have focused on the topic of recovering from moments of miscommunication during speech dialogue in HRI/HAI, which often leads to interaction failures. Typically, that entails some sort of speech about specific future actions on the part of the robot/agent so as to recover from miscommunication by guiding the human user towards a particular behavior [45,46]. While that sort of scenario is still reactive in the sense that it is a response to a previous interaction failure, it does entail some anticipation about the future, though not necessarily utilizing prediction. Rather, for example, speech in that scenario may involve "explanations" of why the current interaction failure may have occurred, so that the human can attempt to fix the issue, e.g., the human performed a task step out of sequence, such as when assembling something [47,48]. That is obviously quite a bit different from speculating about the future. Regardless, it is still an open debate as to how social robots should best respond to failures during HRI, and thus an active area of research in the field [46,49].

Another related topic is that of "fluency" during HAI and HRI, referring to the artificial agent purposely engaging in more proactive behaviors (in contrast to merely reacting). This idea has been explored in various ways by researchers, generally in terms of behaviors intended to trigger human responses [12,50–52]. Some of the early work in that area was done by Hoffman et al., who defined different kinds of robot anticipation and ways to measure any resulting task efficiency improvements [12]. Interestingly, however, later researchers found that proactive behaviors in artificial agents are not always preferable and that, in fact, some users react negatively to such behavior in certain situations [43]. While some of that research has used human speech cues to trigger proactive robot behavior, a substantial portion has utilized non-verbal cues of humans, such as user intent based on eye gaze [50]. Moreover, much of the focus has also been on enabling the robot to interpret human cues, rather than the robot's own behavior [53]. That said, there has also been some related work where the robot attempts to anticipate the human user's next utterance during dialogue, i.e., guessing what the human will say next [54]. This idea is related to the concepts of turn-taking prediction and dialogue state trackers mentioned in Section 1 [21–23]. Nevertheless, there is a significant difference between the body of research on robot fluency and the robot "speculation" we define in Section 1. Similar to the other experimental research above, while there are some things we can draw from the literature for developing our concept of anticipatory speech here, its applicability is limited.

3. Materials and Methods

3.1. Approach

Our approach here incorporates different strategies, combining natural language ontologies with deep learning (DL) models of in-game events (game paradigm described below) [55]. The goal is to predict events likely to occur in the near future and then speak about them *before* they occur. Similar work has been done in prior research using machine learning (ML) for modeling joint actions between humans and robots [56], but our focus here is specifically on conversational aspects (i.e., anticipatory speech) in order to facilitate better interaction. Many of the things we are predicting are game "events" that happen in the world around the human and the artificial agent, not necessarily actions they perform themselves. Nor is there absolute certainty that the events will indeed take place—only the *possibility*. In that sense, it is intended to be similar to how humans communicate with each other via speculation (see Section 1).

3.1.1. Avatar Speech System

We used the Loomie application to create the 3D humanoid virtual avatar for the experiment, using the same ethnically ambiguous female avatar with all participants (https://www.loomielive.com/ accessed on 12 December 2024). The avatar was capable of moving its lips synchronously with the synthesized speech, as well as some basic built-in gestures from the Loomie application, which we did not attempt to modify here. The speech system itself (both development and testing) has been described in detail in several previous publications [57,58], so we provide a brief overview here for brevity. In short, the speech system was developed as a "Social AI" specifically for cooperative game paradigms, where the system has awareness of events occurring within the game in *real time*, including player actions, as well as the ability to detect and respond to human utterances (i.e., automatic speech recognition, ASR). The speech content of the system itself was essentially a knowledge base that comprised hundreds of different speech utterances covering 46 different utterance categories, each related to a particular game situation (e.g., collecting resources, fighting monsters, deciding where to go next).

The agent's speech comprised 3 types: self-generated speech (regarding current/past game events), anticipatory speech (about future possibilities), and ASR responses (to human queries). These are terms we defined to distinguish between utterances the AI produces using its internal logic of what to say next based purely on the current situation (self-generated) versus utterances that are based on predictions of potential future events (anticipatory speech). In contrast to these two, ASR responses are direct responses to human queries. Some short example dialogues of speech interactions between the human user and the artificial agent are provided in Appendix A, including both examples with and without anticipatory speech. The idea was that anticipatory speech should hypothetically create more fluid interactions and provide better coordinated action between the human and the artificial agent.

The anticipatory speech was generated via deep learning models in Keras (https://keras.io/ accessed on 12 December 2024), which have been described in detail previously elsewhere, showing an AUC (area-under curve) performance of roughly 0.81 in predicting future game situations (see Section 4.3 in [58]). Here, we focus on the *application* of those models during user testing and thus provide only a brief description of the model details for brevity. In short, those models attempt to predict future game situations that are likely to occur in the next 5–10 s and then talk about them. The models ran in real time, producing new predictions every second. In terms of architecture, the models combined several layers of convolutional neural networks (CNNs) and recurrent neural networks (LSTM), followed by a final dense layer with a sigmoidal activation function to produce a set of binary

predictions (probabilities) for multiple game situations simultaneously. The idea behind the model architecture was that the CNN layers could parse out "invariant representations" of pattern signatures occurring anywhere in the interaction, followed by the LSTM layers detecting critical "sequences" of those patterns over time. By critical, we mean that those sequences are related to the aforementioned game situations (e.g., collecting resources, fighting monsters, deciding where to go next) occurring shortly thereafter. To limit anticipatory speech to occurring only above a certain predicted probability, a probability threshold was set to control their frequency (0.7 in our case), which was determined through trial and error during testing. If more than one game situation exceeded the threshold, then the game situation with maximum probability was selected. The full detailed results of the training/testing of those models, including a comparison of different model architectures and other types of machine learning models, is reported elsewhere [58], but the deep learning model architecture described above had the best performance (AUC of roughly 0.81). In the study here, we used that optimal model in combination with the aforementioned speech content from our knowledge base, containing 46 different utterance categories, each related to a particular game situation. However, a newer "speech system 2.0" using GPT-3.5 (OpenAI, San Francisco, CA, USA) for on-the-fly utterance generation (rather than a knowledge-based approach) is under development, though it was not available at the time of these experiments.

The speech system was implemented as a custom code written in Python (version 3.10) using locally installed (Windows or Mac) voice packages for text to speech (TTS), with the Microsoft Azure API used for ASR. The speech system is also trilingual, capable of speaking and understanding English, Korean, and Japanese. All of these languages have been rigorously tested and verified with native speakers over several years.

3.1.2. Game Environment

For the study here, we used a video game called "Don't Starve Together" as our cooperative game environment (https://www.klei.com/games/dont-starve-together accessed on 12 December 2024), which anyone can download from Steam online. This video game is a social survival game (e.g., Minecraft) that involves multiple players who need to engage in various tasks to survive as long as possible: collect resources, make tools, fight monsters, find food. Similar to other social survival games, "Don't Starve Together" requires that players collect specific combinations of resources in order to build things, such as weapons and shelters. Lacking those things, they are vulnerable to various dangers and will likely lose the game via player death (i.e., goal-oriented), though there are multiple strategies that can be pursued to achieve those goals (i.e., free-form). More importantly, it has cooperative multi-player gameplay modes (used here), which allow the players to cooperate on such tasks to survive. The tasks are under time constraints, however, since the danger level and difficulty gradually increase over time. Given those attributes, "Don't Starve Together" provides an ideal platform to simulate free-form yet goal-oriented cooperative gameplay that mirrors the one found in real-world interaction scenarios, where there are typically multiple possible strategies and time constraints on actions/decisions.

3.2. Experiment and Data Collection

For the experiment here, we recruited 40 participants, with 20 English speakers and 20 Korean speakers. We then randomly split the participants up, stratified by language, so that 20 of them were assigned the experimental H1 condition (with "anticipatory speech" enabled) and 20 were assigned a Control condition (without it, i.e., disabled). The result was that each condition had 10 participants in Korean and 10 in English, which allowed us to test for difference by condition and by language simultaneously. The study design and

sample size were based on effect sizes (>0.8) observed in our previous studies of bilingual robots and were approved by the IRB of Hanyang University (protocol # HYU-2021-138).

An example of the experiment can be seen in Figure 1. The experiment's physical setup comprised two computers in two separate rooms, with one for the human participant ("player computer") and one for the virtual avatar where its code was run ("confederate computer"). Both computers were linked to the same online game server. The player computer was further equipped with an HD camera, headphones, and Blue Snowball microphone for high-quality audiovisual recordings of the experiment. Each game session involved one human participant and the virtual avatar interacting autonomously during a 30 min game session on a private server in a 2-player cooperative gameplay mode. The two could see and communicate with each other via Zoom during the entire game session, alongside the game itself.

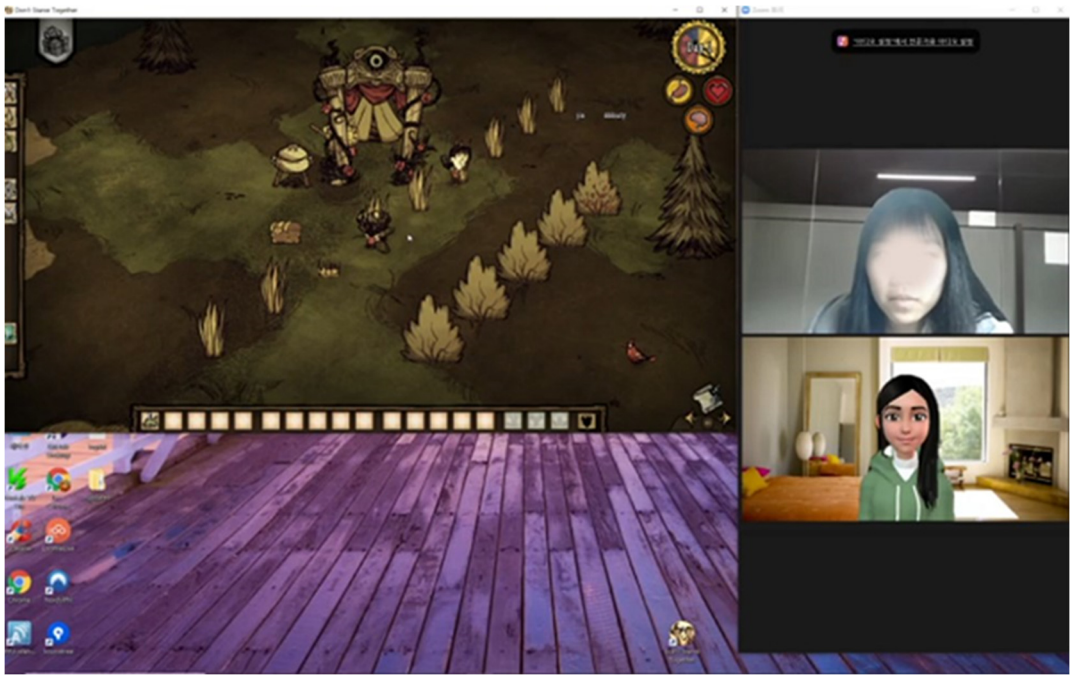

Figure 1. Gameplay example during experiment: human vs. avatar (from [58]).

3.3. Data Analysis

During the experiments, we collected audiovisual recordings of the interaction between the human and the virtual avatar during each game session using OBS Studio (https://obsproject.com/ accessed on 12 December 2024). From those recordings, we later extracted the speech for both the avatar and the human player synced with in-game gameplay events, which were then used for various kinds of natural language processing (NLP) analysis to generate speech interaction patterns. Speaker diarization (available open-source from Google) was used to determine whether the human or avatar was speaking at any given moment, while standard NLP methods were used to generate utterance counts, interruption frequencies (based on interpausal units, IPUs [21]), and speech sentiment (lexical parsing via VADER [59,60]), similar to our previous publications [57,61]. These measures are typically used to assess the quality of the speech interaction. We utilized two-tailed independent-sample t-tests on those speech interaction patterns to test for differences across conditions and languages.

We also collected several standardized HRI instruments at the end of the experiment for each participant, including Godspeed [62] to measure human perceptions of the robot/agent and Networked Minds [63] to measure "social presence" [64]. All of

these instruments were evaluated across languages and conditions, the same as the speech interaction patterns above.

4. Results

We evaluated several different metrics meant to be indicators of the "quality" of the speech interaction with artificial agents, including utterance counts, interruption frequencies, and speech sentiment, similar to other research in the field [21,57,59,61]. These were designed to allow us to get a sense of aspects such as how engaging a conversation was (utterance counts) and how fluid it was (based on interruptions). We also know from past research that these metrics vary by language [57].

4.1. Utterance Counts

First, we wanted to examine the amount of speech (i.e., utterance counts) to see if there were any effects due to adding anticipatory speech capabilities to an artificial agent. Before doing so, we checked the speech output of the agent in the H1 condition to confirm if it was producing a reasonable number of anticipatory speech utterances while still producing its normal self-generated speech (on current/past game events) and ASR responses (to human queries). As shown in Figure 2, that appeared to be the case, with anticipatory speech comprising about 30% of the agent's total speech (the original design intention).

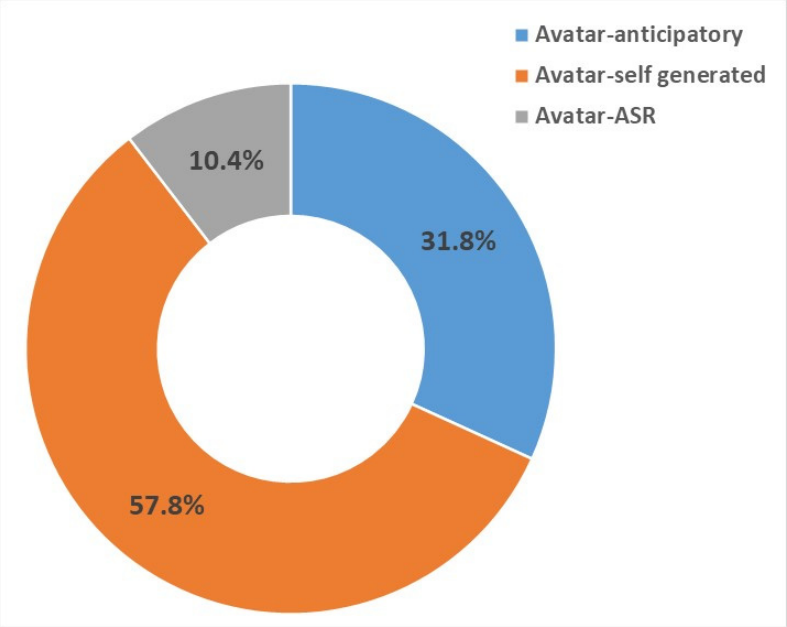

Figure 2. Artificial agent speech patterns (H1 condition).

After validating that the system was working as designed, we analyzed the average amount of speech utterances for both the human and agent across game sessions, by both language and condition. The results can be seen in Table 1. The main takeaway was that introducing anticipatory speech in the artificial agent (H1 condition) caused both the human and the agent to talk more than without such speech (Control condition), including the agent producing more self-generated speech based on game events. However, those differences depended on the language being spoken, with the significant differences mainly being on the Korean side. In the Control condition, Korean speakers spoke much less with the agent than English speakers (or bilingual speakers, for that matter), which we have discussed in previous publications [57,61]. However, with anticipatory speech added, Korean speakers were much more similar to English speakers. We do note there was an

increase in English human speakers' utterances on average, though that did not obtain statistical significance here, but might if the study was re-run with a larger sample size.

Table 1. Utterance counts by condition and language (* <0.05, ** <0.01, *** <0.001).

	Control (std)	H1 (std)	*p*-Value	*Sign.*
English				
Human	122.66 (54.8)	169.9 (79.29)	0.15390	
Avatar	99.44 (18.02)	135 (34.37)	0.01250	*
Avatar-planned	NA	32.8 (8.92)	-	
Avatar-self generated	79.33 (10.93)	80.5 (20.24)	0.87640	
Avatar-ASR	20.11 (11.24)	21.7 (13.63)	0.78640	
Korean				
Human	35.1 (18.46)	96.8 (52.35)	0.00480	**
Avatar	32.6 (17.68)	119 (23.88)	0.00000	***
Avatar-planned	NA	48 (15.28)	-	
Avatar-self generated	30.8 (16.87)	66.2 (10.05)	0.00002	***
Avatar-ASR	1.8 (1.48)	4.8 (4.05)	0.04990	*

Curiously, the increase in the artificial agent's self-generated speech here also suggests there may have been a change in the human playing the game, perhaps more adventurously, leading to more game events that the artificial agent could comment about. In other words, anticipatory speech might possibly affect human non-verbal behavior during HRI/HAI. This is an intriguing possibility, though our attempts to confirm this hypothesis by analyzing the gameplay data produced only mixed results. More research (and/or possibly a different study design) is needed to tease that apart. As such, it remains an intriguing but speculative idea for now.

4.2. Interruption Frequency

Next, we evaluated whether adding anticipatory speech caused any changes in how frequently either the human interrupted the avatar, or vice versa. Interruptions here are defined as speech that overlapped the other player's speech (IPU), in contrast to waiting for them to finish speaking, regardless of whether it was accidental or not. In a broader sense, we can think of such interruptions as failures of proper turn-taking during a given social interaction [21]. Since there were differences in the utterance counts between language and condition (see Section 4.1), the interruption counts were calculated as a percentage of the total utterance count within each category, for fair comparison. In other words, they are frequencies based on the number of "opportunities to interrupt", rather than raw counts. The results can be seen in Table 2, broken out by language and by which speaker interrupted the other.

Table 2. Interruption frequency by language and speaker (** <0.01, *** <0.001).

	Control (std)	H1 (std)	*p*-Value	*Sign.*
Condition				
English	2.96% (3.30)	3.70% (2.10)	0.40280	
Korean	0.76% (1.29)	5.16% (3.89)	0.00010	***
Overall	1.80% (2.77)	4.43% (3.21)	0.00850	**
Speaker				
Avatar—English	4.20% (3.09)	4.56% (1.78)	0.65420	
Human—English	1.68% (1.33)	2.85% (2.25)	0.05250	
Avatar—Korean	0.90% (1.57)	8.09% (3.40)	0.00010	***
Human—Korean	0.64% (0.90)	2.23% (1.73)	0.00080	***

Similar to what we saw in Section 4.1, for interruption frequency, all significant differences were on the Korean side, with none on the English side. The largest difference by far was with the Korean-speaking virtual avatar agent (Control—0.9% vs. H1—8.1%), which had a roughly 8-fold increase in interruption frequency. Meanwhile, on the English side, the agent's interruption frequency was virtually the same. There are a number of possible reasons why that may have occurred. For instance, perhaps English speakers, by nature, expect more anticipatory speech to begin with, so they leave more pauses during a conversation to allow for that. If so, that could explain why there was less of an effect of anticipatory speech on the English side for utterance counts in Section 4.1. This could also explain why the Korean speakers talked more during the H1 condition, if perhaps the unexpected anticipatory speech prompted in them an increased urge to respond to the agent. Perhaps it even reflects a cultural/linguistic difference in verbal planning behavior, which might be embedded in the structure of the language itself [65].

However, the above is just one possible interpretation that we considered, and there could be other interpretations. For instance, it could be related to differences in politeness levels across cultures. Alternatively, perhaps there are some non-verbal cues to turn-taking that are important on the Korean side, but not so much on the English side, which we unknowingly omitted due to the virtual nature of the interaction. That is the topic we return to in Section 5.2 in the Section 5, where we discuss the design implications for artificial agents.

4.3. Other Analysis Results

We also conducted several other analyses to test for other effects due to anticipatory speech in artificial agents, none of which obtained statistical significance. We report those here, for thoroughness.

We analyzed the speech sentiment by both language and speaker, comparing the Control and H1 condition. The results are shown in Figures 3 and 4. Although there was a slight shift towards more positive/neutral sentiment and less negative sentiment (in both Korean and English) when the artificial agent employed anticipatory speech during the H1 condition, those shifts were not statistically significant in our study here.

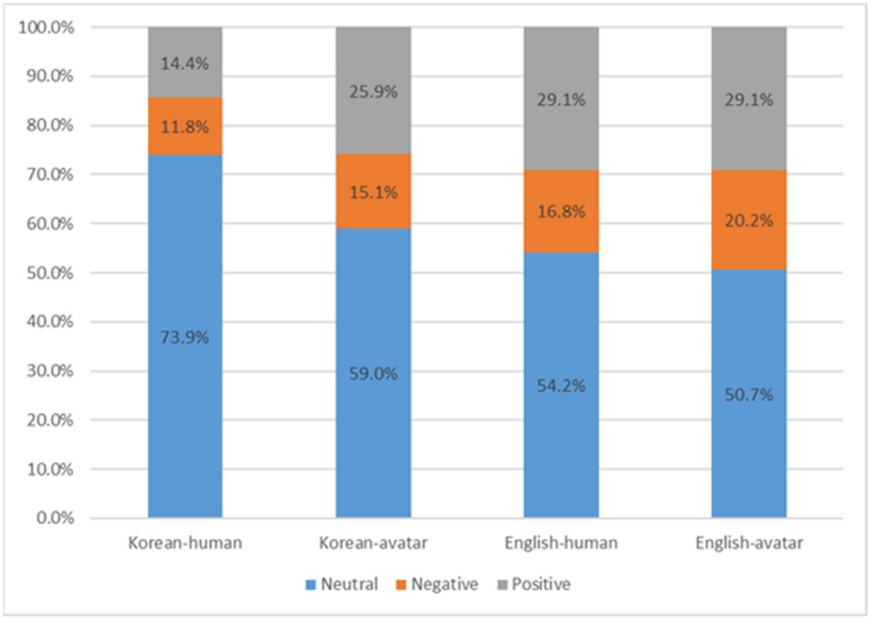

Figure 3. Sentiment analysis—H1 condition.

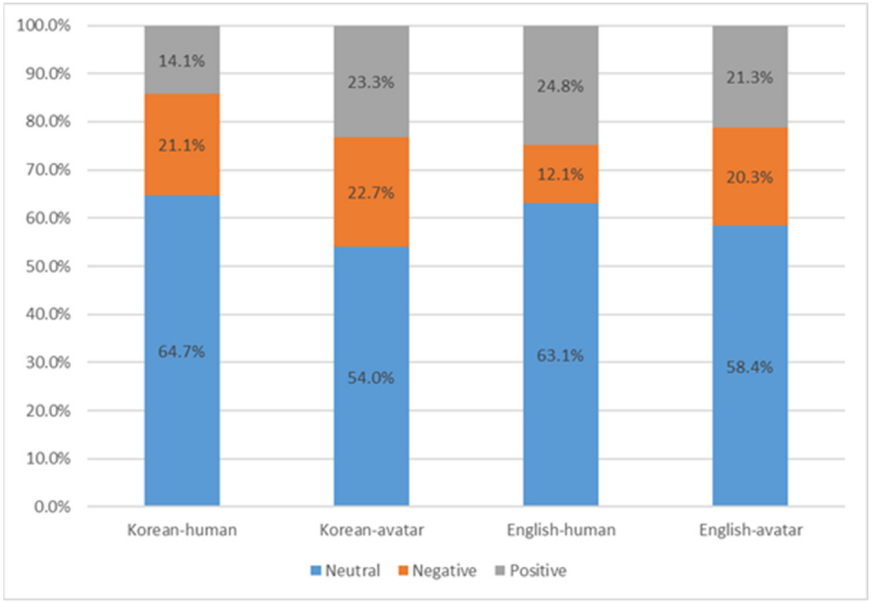

Figure 4. Sentiment analysis—Control condition.

Likewise, we analyzed the collected HRI instrument data mentioned in Section 3.3. The results can be seen in Table 3. There were no significant differences in either Godspeed or Network Minds with and without anticipatory speech. Nor were there any differences separately for Korean or English, or for any of the individual subscales from Godspeed. This implies that anticipatory speech does not significantly affect human perceptions of the artificial agent or the sense of "social presence"—at least not in a way detectable by the instruments used in the study design here. We cannot rule out that other HRI instruments may be more sensitive to the effects of anticipatory speech on user perceptions. More research is needed on the topic, as presumably, if there is a behavioral effect, then there must be some perceptual effect (even if at only a subconscious level).

Table 3. Instrument comparison by condition.

	Control	H1	*p*-Value	*Sign.*
Godspeed	3.44	3.22	0.16100	
NM Self	3.20	3.07	0.27000	
NM Other	3.31	3.09	0.20200	

5. Discussion

5.1. Summary of Results

Based on our results here, it appears to be clear that adding anticipatory speech to artificial agents and robots does affect the way in which people communicate with them. These effects on the interaction, however, appear to vary by language and context. In other words, what we find during an English-based HRI/HAI study may really only apply to English, rather than some universal principle across all languages and/or contexts. Beyond that, we also detected some potential changes in the humans' behavior due to anticipatory speech. That included both altered speech behaviors and in-game behavior, though the results for the latter were not conclusive and require more study.

To summarize, there are clear differences during human interaction when artificial agents engage in anticipatory speech, speculating about future events. The effects thereof appear to play out in some unexpected ways, which vary based on a number of factors. This is a rich area for future research in HRI/HAI to explore, and it also has a number of

implications for how we design robots and artificial agents, as well as for how we think about modes of communication during interaction. We explore these implications in the next section.

5.2. Implications

5.2.1. Design Implications for Artificial Conversational Agents

There are a number of implications from this research for both designing artificial conversational agents and thinking about their use in real-world in-home environments. For instance, in this study, we saw that alterations in the speech system of such agents can have differing effects across languages, aligning with results seen in other studies [57,61]. This is critical to consider in the modern world, where multi-cultural and multi-lingual environments are increasingly common [8,66,67]. In short, a CUI designed for a multi-cultural and/or multi-lingual environment may need to be designed differently from a more homogenous monolingual one.

We can extend the above point more broadly to think about how context affects interactions with conversational agents, which is unfortunately too often overlooked [68]. Such a context may involve culture and linguistic setting, but also a human's functional role at the current moment. For instance, at work, we may serve one role (e.g., bank teller, nurse, professor), while serving a different role at home (e.g., parent, caregiver, spouse). The exact same person can behave (and communicate) in very different ways depending on the context and their role within it. Moreover, those contexts can change depending on our location and/or who else is present at the time. This demands that we consider *the socially situated context* within which an interaction takes place [67].

For conversational studies, such as the one we conducted here, the above will likely necessitate coming up with some way to disentangle the lexical aspects of a language itself (i.e., the words used) from the cultural influences that shape how individual native speakers express themselves through those words. Those lexical aspects are, of course, rooted in the historical development of the language, developing alongside the culture over time. Meanwhile, culture influences how individuals behave/speak in the present, which may diverge from the historical lexicon for various reasons [7]. For example, think of the difference between the standard meaning of a word and its use as slang, e.g., the word "snack" in modern American English (referring to an attractive person, rather than food). Disentangling the two will be no easy feat and likely require a very sophisticated study design. One possibility may be to take advantage of slang meanings that diverge from standard meanings to investigate *socially situated context* effects on conversations between humans and robots/agents, though how to do so cross-linguistically via experiments will be a challenge.

Another design aspect is to think about the technological context. In short, the conversational agent (or the device it is speaking through) is never going to be the only technology in the room. Indeed, with the growth of internet-of-things (IOT) systems in homes and workplaces, humans are increasingly living in a "world of devices" [69]. One thing to consider is how a conversational agent on some specific device can interface with other devices in the environment or leverage data coming off such devices. For example, in a healthcare setting, there may be many digital health devices present (e.g., wearables), which could all provide a different perspective on the patient's status through different sensing capabilities [70]. That may allow the conversational agent to autonomously alter its dialogue in ways that better assist a patient, taking into account linguistic differences in how individuals express their physical/mental state or incorporating "speech biomarkers" from the conversation itself [22]. Or take a different scenario. Imagine a conversational physical robot that could "transform" into an artificial agent and appear on a digital screen

elsewhere in a person's home in order to continue communication—or even transfer to a digital screen in the person's vehicle when they leave home. It is important for the field of human-centered AI to think about these kinds of future interaction scenarios [71,72]. It may not be necessary to have a physical device that follows the person from room to room or to different locations, if devices already exist in those places. That is a radically different way to think about "robots".

5.2.2. Design Implications for HRI/HAI Gameplay

We would be remiss not to note a few design considerations specifically for human gameplay with artificial agents and robots based on the experiments described in this paper. There are two major things of note.

First, during a computer vision (CV) analysis of human facial expressions, we discovered that during these kinds of cooperative video game scenarios, human users tended to make limited eye contact with the virtual avatar—only on occasion and often very briefly. This is in contrast to other previously reported HRI experiments [73–75], which we suspect has something to do with the intensive rapid-fire pace of the game used here. Many of those previously reported games in HRI experiments were slow-paced turn-taking kinds of games where players can pause and think about their next move (~70–80% of experiments use such games [76]), but modern video games are often much more fast-paced and require sustained attention.

Second, we found that human users often make only a small range of facial expressions during these games, typically alternating between either a happy expression (smiling or laughing) and a look of concentration or frustration, which modern facial expression recognition algorithms tend to classify as "neutral". Other facial expressions (e.g., sadness, surprise, disgust) were recorded very rarely. We have reported that CV analysis elsewhere in a separate journal publication [77].

There are some important design implications for HRI that we can draw from the two points above. In particular, the types of interactions we observe between humans and robots during gameplay will likely depend on the pace of the game and the level of attention required. which appears to be true even for cooperative games, so we should be careful generalizing conclusions only from a single type of game, as has often been done in previous research. Such caution has broad applicability, given the growing interest in "gamification" for human-centered AI systems [78].

5.2.3. Physical vs. Virtual Interaction in HRI

There is ongoing debate in the scientific community about the difference between virtual interaction and physical interaction, not only in HRI but also in fields such as HAI, CUI, etc. Nevertheless, there is also increasing interest within those fields in "mixed-reality" platforms that combine the virtual and physical, as well as virtual simulations of physical interactions for design purposes [27,79,80]. Critically, these approaches can allow us to more rapidly prototype social interaction scenarios. Such interest was underscored by the recent, first-ever CUI workshop at the ACM HRI flagship conference in 2023, which drew a large audience from multiple overlapping scientific fields [81].

One aspect that cannot be ignored, especially when creating conversational agents and user interfaces that appear through a screen rather than a physical robot (as we did here), is the importance of non-verbal social cues [82,83]. This obviously leads to the question of how virtual interaction paradigms can be utilized to design such physical social cues. There are some examples of this—e.g., Bartneck et al. (2015) looked at using Unity to visualize and create more realistic movements for social robots [84]. Other researchers have studied the use of physical haptic feedback as a social cue while users wear a virtual reality

(VR) headset during a virtual interaction with a physical robot [85], while others have examined applying general UX design principles of digital interfaces to reimagine physical interactions as a kind of step-by-step "interface" [86]. Another common area of research is using VR systems for multi-robot control in human-in-the-loop scenarios, where otherwise coordinating the actions of multiple robots is quite difficult for a human user [87].

There are also many examples of using virtual interaction as a substitute for physical interaction during human–robot communication. A common problem in HRI is having our robot indicate its intent to human users somehow through mainly physical means (e.g., gaze), leading to some ambiguity that often leaves the human guessing [88]. That can sometimes be augmented through verbal cues towards specific objects, though the scalability of such systems remains unclear. However, another avenue is to provide so-called "mixed reality" interfaces combining virtual reality and other *digital interfaces* into the system. We are already seeing some HRI platforms adopt such an approach. For instance, many robots now include digital interfaces (e.g., screens) that allow us to incorporate both embodied interaction and virtual interaction on the same platform [89–91]. A great example of the above is the Furhat, which is a 3D digital image of a face retro-projected from behind onto a plastic screen made in the shape of a human face, but a virtual image nonetheless [92,93].

5.3. Limitations

There are a number of limitations to this research, both theoretical and technical, of which we note a few specific ones here. First, one major limitation is the use of the English language for comparison. Many studies in psycho-linguistics and cognitive science focus on comparisons of English versus some second language, but obviously, the linguistic world is much broader than that, and there may be idiosyncrasies about the English language that are not representative of all languages [94]. Moreover, given the global popularity of American pop culture, often even monolingual speakers of another language have likely been exposed to at least a small amount of English during their lifetime, which is problematic. Obviously, there are logistical constraints on the number of languages for which we have access to native L1 speakers in any one geographical location, which was part of the reason the study here was limited to two languages (Korean and English). Thus, if we want to expand the number of languages we compare in HRI/HAI research, it will be necessary to find new ways of doing research, perhaps by using international collaborative networks of researchers in different locations. We discuss this issue at length in another paper [22].

Second, there are potentially significant differences between virtual agents and physical agents when it comes to human interaction, though to what degree is still subject to debate (see Section 5.2). This topic has been extensively studied in the field of HRI, including some of our own past work [95,96]. Nevertheless, the experiments described here should be replicated in the future on a variety of HRI/HAI platforms, both physical and digital, to compare the consistency of the results. However, there are some technical challenges in transferring knowledge gained from virtual experiments onto physical robot platforms, which has been discussed elsewhere [97]. Even simple things like physical lip-syncing during speech by a robotic face can represent significant engineering challenges.

Third, we note that although the sample sizes used here were judged sufficient as they were based on the rather large effect sizes (>0.8) observed in our previous studies on bilingual human–agent conversation [49,57,58,61], there is a need to replicate these findings on bigger sample sizes for further verification. This exploratory study can hopefully serve as the groundwork for that from a methodological standpoint. Likewise, some of our suggestions from Section 5.2.1, such as the comparison of slang use and standard

use of language, could be investigated during those larger experiments to explore the *socially situated context* of how language is used during human verbal behavior in order to disentangle the interaction of language (i.e., lexicon) and culture.

Finally, there needs to be more exploration into how the appearance of an avatar/robot might affect speech interactions. There has been some work on this aspect by other researchers, though not in the exact same context as the experiments here [98,99]. We did not consider that in this research study because the appearance of the virtual avatar was purposely fixed so that we could evaluate other factors, though we are currently running another set of experiments that do vary the physical appearance. Suffice it to say, additional research is needed.

Author Contributions: Conceptualization, C.C.B., S.Y.K. and B.W.; Methodology, C.C.B., S.Y.K. and B.W.; Software, Y.-H.B. and J.-H.Y.; Formal analysis, C.C.B., Y.-H.B. and J.-H.Y.; Investigation, C.C.B., Y.-H.B. and J.-H.Y.; Writing—original draft, C.C.B.; Writing—review & editing, Y.-H.B. and J.-H.Y.; Funding acquisition, C.C.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a grant from the National Research Foundation of Korea (NRF), under grant number 2021R1G1A1003801.

Data Availability Statement: Data used in this study may be available in a de-identified form upon request from the authors.

Acknowledgments: We thank the anonymous reviewers for their valuable comments about this paper. We would also like to thank all of our undergraduate research assistants (Hanyang University) for their assistance in this work.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

This appendix includes several examples of dialogues between the human user and the artificial agent, including both with and without anticipatory speech. Anticipatory speech utterances are highlighted in blue, with the predicted scenario indicated in parentheses in the table caption (when applicable). The dialogues encompass short snippets covering about 10–15 s.

Anticipatory?	Speaker	Utterance
Yes	Avatar	We need a place to store resources.
Yes	Avatar	You should gather resources around you.
	Player	OK, let's go a little more in the southern
	Player	Get some twigs
	Avatar	Guess I can make an axe now!
	Avatar	Let's fight!
	Player	Alright, I'll be careful
	Avatar	Crap, I'm being attacked by this (killer bee)!
	Player	I'm coming to you
	Player	Well you have already crushed it

Figure A1. Speech dialogue with anticipatory speech (gathering resources).

Anticipatory?	Speaker	Utterance
	Player	Do you have some food?
	Player	My health is low so i need to eat something
	Avatar	Gonna attack that (spider)!
	Player	Yeah, we should not attack him right now
	Player	I might have something to attack them later
	Avatar	Let's finish this off quickly together!
	Avatar	Aaah, my health is critical!
	Avatar	Help me fight this thing!
	Player	Yeah, we should make a campfire (so the monster leaves)
	Player	Ohh no no i cannot, I need some more grass
	Avatar	Run away!

Figure A2. Speech dialogue without anticipatory speech.

Anticipatory?	Speaker	Utterance
Yes	Avatar	Think there's a monster around. Be careful
Yes	Avatar	Can you make me an axe?
	Player	Yes, hold on
	Player	Let me give you this axe, here
	Player	Have some petals too
	Avatar	Think I have plenty of food.
	Player	The tank. That guy. Ohh
	Player	Ohh this thing
	Avatar	Let's get that (spider den)!
	Avatar	Help me fight this thing!
	Avatar	Dammit! This (spider) is attacking me.
	Player	Ok, wait a minute. let me help you. let me help

Figure A3. Speech dialogue with anticipatory speech (fighting monsters).

Anticipatory?	Speaker	Utterance
	Avatar	Let's hunker down for the night.
	Player	My god, but how to remove the fire?
	Player	How can you sleep?
	Avatar	That's better, I hate darkness
	Player	Well, there is fire everywhere
	Avatar	Least it's not pitch black.
	Player	Ah, but there is fire everywhere
	Avatar	Feel better now, it is scary when it is dark.
	Player	Ok, let me sleep
	Player	I think I should have had something (to eat)
	Avatar	I need to get more food!
	Player	Fine, me too
	Avatar	Let's find some soon
	Player	Where, where can we find food?

Figure A4. Speech dialogue without anticipatory speech.

Anticipatory?	Speaker	Utterance
	Player	Let's go back, to our home
Yes	Avatar	I found a good place to rest. We could rest here
	Avatar	Feel better now, it is scary when it is dark.
	Player	Let's just cook up these eggs here
	Player	Can you give me some grass and twigs? Thank you
	Player	Oh god. We're going insane
	Player	I can see that
	Avatar	We should get some sleep.
	Avatar	Let's cook food and make up for our hunger.
	Player	Sounds good

Figure A5. Speech dialogue with anticipatory speech (impending night-time, i.e., substantial danger level increase).

Anticipatory?	Speaker	Utterance
	Player	Okay, i'll give you some petals
	Player	I agree, let's use the science machine
	Player	First have to create a science machine
	Player	Oh, we need gold for that
	Player	Okay, so tomorrow we'll look for golds
	Avatar	Ah, another night another day.
	Avatar	Let's go this way!
	Player	Let's find some gold
	Player	Will you slow down? Oh nice (look over there)
	Avatar	Let's fight! Gonna attack that (spider)!
	Avatar	Help me out! Aaaaah!
	Player	Later. But you're fine
	Player	Let's find some gold first

Figure A6. Speech dialogue without anticipatory speech.

References

1. Byrne, R.M.J. *The Rational Imagination: How People Create Alternatives to Reality*; MIT Press: Cambridge, MA, USA, 2007.
2. Atance, C.M.; O'Neill, D.K. The emergence of episodic future thinking in humans. *Learn. Motiv.* **2005**, *36*, 126–144. [CrossRef]
3. Li, H.H.; Curtis, C.E. Neural population dynamics of human working memory. *Curr. Biol.* **2023**, *33*, 3775–3784. [CrossRef] [PubMed]
4. Abrams, A.M.; der Pütten, A.M.R.V. I–C–E Framework: Concepts for Group Dynamics Research in Human-Robot Interaction: Revisiting Theory from Social Psychology on Ingroup Identification (I), Cohesion (C) and Entitativity (E). *Int. J. Soc. Robot.* **2020**, *12*, 1213–1229. [CrossRef]
5. Kantharaju, R.B.; Pelachaud, C. Social Signals of Cohesion in Multi-party Interactions. In Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, Virtual, 14–17 September 2021; pp. 9–16. [CrossRef]
6. Pineda, J.A. Is Social Cohesion a Different Mechanism of Evolution? In *The Social Impulse: The Evolution and Neuroscience of What Brings Us Together*; Springer International Publishing: Cham, Switzerland, 2021; pp. 43–51. [CrossRef]
7. Bennett, C.C.; Lee, M. Would People Mumble Rap to Alexa? In Proceedings of the 5th International Conference on Conversational User Interfaces (CUI), Eindhoven, The Netherlands, 19–21 July 2023; pp. 1–5. [CrossRef]
8. Lim, V.; Rooksby, M.; Cross, E.S. Social robots on a global stage: Establishing a role for culture during human–robot interaction. *Int. J. Soc. Robot.* **2021**, *13*, 1307–1333. [CrossRef]
9. Marge, M.; Espy-Wilson, C.; Ward, N.G.; Alwan, A.; Artzi, Y.; Bansal, M.; Blankenship, G.; Chai, J.; Daumé, H.; Dey, D.; et al. Spoken language interaction with robots: Recommendations for future research. *Comput. Speech Lang.* **2022**, *71*, 101255. [CrossRef]

10. Lallée, S.; Hamann, K.; Steinwender, J.; Warneken, F.; Martienz, U.; Barron-Gonzales, H.; Pattacini, U.; Gori, I.; Petit, M.; Metta, G.; et al. Cooperative human robot interaction systems: IV. Communication of shared plans with Naïve humans using gaze and speech. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Tokyo, Japan, 3–7 November 2013; pp. 129–136. [CrossRef]
11. Talamadupula, K.; Briggs, G.; Chakraborti, T.; Scheutz, M.; Kambhampati, S. Coordination in human-robot teams using mental modeling and plan recognition. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Chicago, IL, USA, 14–18 September 2014; pp. 2957–2962. [CrossRef]
12. Hoffman, G. Anticipation in human-robot interaction. In Proceedings of the AAAI Spring Symposium Series, Palo Alto, CA, USA, 22–24 March 2010.
13. Briggs, G.; Scheutz, M. The pragmatic social robot: Toward socially-sensitive utterance generation in human-robot interactions. In Proceedings of the AAAI Fall Symposium Series, Arlington, VA, USA, 17–19 November 2016.
14. Castellucci, G.A.; Kovach, C.K.; Howard, M.A., III; Greenlee, J.D.; Long, M.A. A speech planning network for interactive language use. *Nature* **2022**, *602*, 117–122. [CrossRef]
15. Postma, A.; Kolk, H.; Povel, D.J. Speech planning and execution in stutterers. *J. Fluen. Disord.* **1990**, *15*, 49–59. [CrossRef]
16. Borthwick, S. Communication impairment in patients following stroke. *Nurs. Stand.* **2012**, *26*, 35. [CrossRef]
17. Irfan, B.; Hellou, M.; Belpaeme, T. Coffee with a hint of data: Towards using data-driven approaches in personalised long-term interactions. *Front. Robot. AI* **2021**, *8*, 676814. [CrossRef]
18. Rastogi, A.; Hakkani-Tür, D.; Heck, L. Scalable multi-domain dialogue state tracking. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 561–568. [CrossRef]
19. Lim, J.; Whang, T.; Lee, D.; Lim, H. Adaptive Multi-Domain Dialogue State Tracking on Spoken Conversations. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *32*, 727–732. [CrossRef]
20. Bae, Y.H.; Bennett, C.C. Real-Time Multimodal Turn-taking Prediction to Enhance Cooperative Dialogue during Human-Agent Interaction. In Proceedings of the 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Busan, Republic of Korea, 28–31 August 2023; pp. 2037–2044. [CrossRef]
21. Skantze, G. Turn-taking in conversational systems and human-robot interaction: A review. *Comput. Speech Lang.* **2021**, *67*, 101178. [CrossRef]
22. Bennett, C.C. Findings from Studies on English-Based Conversational AI Agents (including ChatGPT) Are Not Universal. In Proceedings of the 6th ACM Conference on Conversational User Interfaces (CUI), Luxembourg, 8–10 July 2024; pp. 1–5. [CrossRef]
23. Irfan., B.; Kuoppamäki, S.M.; Skantze, G. Between reality and delusion: Challenges of applying large language models to companion robots for open-domain dialogues with older adults. *ResearchSquare PrePrint* **2023**. [CrossRef]
24. Nordberg, O.E.; Guribye, F. Conversations with the News: Co-speculation into Conversational Interactions with News Content. In Proceedings of the 5th International Conference on Conversational User Interfaces (CUI), Eindhoven, The Netherlands, 19–21 July 2023; pp. 1–11. [CrossRef]
25. Lee, M.; Noortman, R.; Zaga, C.; Starke, A.; Huisman, G.; Andersen, K. Conversational futures: Emancipating conversational interactions for futures worth wanting. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI), Yokohama, Japan, 8–13 May 2021; pp. 1–13. [CrossRef]
26. Aylett, M.P.; Cowan, B.R.; Clark, L. Siri, echo and performance: You have to suffer darling. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI), Glasgow, UK, 4–9 May 2019; pp. 1–10. [CrossRef]
27. Williams, T.; Szafir, D.; Chakraborti, T.; Ben Amor, H. Virtual, augmented, and mixed reality for human-robot interaction. In Proceedings of the Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI), Chicago, IL, USA, 5–8 March 2018; pp. 403–404. [CrossRef]
28. Suzuki, R.; Karim, A.; Xia, T.; Hedayati, H.; Marquardt, N. Augmented reality and robotics: A survey and taxonomy for AR-enhanced human-robot interaction and robotic interfaces. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI), New Orleans, LA, USA, 29 April–5 May 2022; pp. 1–33. [CrossRef]
29. Völkel, S.T.; Schneegass, C.; Eiband, M.; Buschek, D. What is" intelligent" in intelligent user interfaces? a meta-analysis of 25 years of IUI. In Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI), Cagliari, Italy, 17–20 March 2020; pp. 477–487. [CrossRef]
30. Brooks, R.A. Elephants don't play chess. *Robot. Auton. Syst.* **1990**, *6*, 3–15. [CrossRef]
31. Pezzulo, G. Anticipation and future-oriented capabilities in natural and artificial cognition. In *50 Years of Artificial Intelligence: Essays Dedicated to the 50th Anniversary of Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 257–270.
32. Chater, N. How could we make a social robot? A virtual bargaining approach. *Philos. Trans. R. Soc. A* **2023**, *381*, 20220040. [CrossRef] [PubMed]
33. Sagar, M.; Henderson, A.M.E.; Takac, M.; Morrison, S.; Knott, A.; Moser, A.; Yeh, W.-T.; Pages, N.; Jawed, K. Deconstructing and reconstructing turn-taking in caregiver-infant interactions: A platform for embodied models of early cooperation. *J. R. Soc. N. Z.* **2023**, *53*, 148–168. [CrossRef]

34. Knott, A.; Sagar, M.; Takac, M. The ethics of interaction with neurorobotic agents: A case study with BabyX. *AI Ethics* **2022**, *2*, 115–128. [CrossRef]
35. Gross, S.; Krenn, B. A communicative perspective on human–robot collaboration in industry: Mapping communicative modes on collaborative scenarios. *Int. J. Soc. Robot.* **2024**, *16*, 1315–1332. [CrossRef]
36. Anastasopoulou, I.; van Lieshout, P.; Cheyne, D.O.; Johnson, B.W. Speech Kinematics and Coordination Measured With an MEG-Compatible Speech Tracking System. *Front. Neurol.* **2022**, *13*, 828237. [CrossRef]
37. Martins, G.S.; Santos, L.; Dias, J. User-adaptive interaction in social robots: A survey focusing on non-physical interaction. *Int. J. Soc. Robot.* **2019**, *11*, 185–205. [CrossRef]
38. Williams, T.; Thames, D.; Novakoff, J.; Scheutz, M. Thank you for sharing that interesting fact! Effects of capability and context on indirect speech act use in task-based human-robot dialogue. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI), Chicago, IL, USA, 5–8 March 2018; pp. 298–306. [CrossRef]
39. Law, T.; Scheutz, M. Trust: Recent concepts and evaluations in human-robot interaction. In *Trust in Human-Robot Interaction*; Academic Press: Cambridge, MA, USA, 2021; pp. 27–57. [CrossRef]
40. St. Clair, A.; Mataric, M. How robot verbal feedback can improve team performance in human-robot task collaborations. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI), Portland, OR, USA, 2–5 March 2015; pp. 213–220. [CrossRef]
41. Grover, T.; Rowan, K.; Suh, J.; McDuff, D.; Czerwinski, M. Design and evaluation of intelligent agent prototypes for assistance with focus and productivity at work. In Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI), Cagliari Italy, 17–20 March 2020; pp. 390–400. [CrossRef]
42. Kraus, M.; Wagner, N.; Minker, W. Effects of proactive dialogue strategies on human-computer trust. In Proceedings of the 28th ACM conference on User Modeling, Adaptation and Personalization, Genoa, Italy, 14–17 July 2020; pp. 107–116. [CrossRef]
43. Xie, L.; Liu, C.; Li, D. Proactivity or passivity? An investigation of the effect of service robots' proactive behaviour on customer co-creation intention. *Int. J. Hosp. Manag.* **2022**, *106*, 103271. [CrossRef]
44. Tian, L.; Oviatt, S. A taxonomy of social errors in human-robot interaction. *ACM Trans. Hum. Robot Interact. (THRI)* **2021**, *10*, 1–32. [CrossRef]
45. Marge, M.; Rudnicky, A.I. Miscommunication detection and recovery in situated human–robot dialogue. *ACM Trans. Interact. Intell. Syst. (TiiS)* **2019**, *9*, 1–40. [CrossRef]
46. Honig, S.; Oron-Gilad, T. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Front. Psychol.* **2018**, *9*, 861. [CrossRef] [PubMed]
47. Mok, B.K.J.; Yang, S.; Sirkin, D.; Ju, W. A place for every tool and every tool in its place: Performing collaborative tasks with interactive robotic drawers. In Proceedings of the 24th IEEE international symposium on robot and human interactive communication (RO-MAN), Kobe, Japan, 31 August–4 September 2015; pp. 700–706. [CrossRef]
48. Das, D.; Banerjee, S.; Chernova, S. Explainable AI for robot failures: Generating explanations that improve user assistance in fault recovery. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI), Boulder, CO, USA, 8–11 March 2021; pp. 351–360. [CrossRef]
49. Bennett, C.C.; Weiss, B. Purposeful failures as a form of culturally-appropriate intelligent disobedience during human-robot social interaction. In Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS): Best & Visionary Papers, Virtual, 9–13 May 2022; pp. 84–90. [CrossRef]
50. Newman, B.A.; Biswas, A.; Ahuja, S.; Girdhar, S.; Kitani, K.K.; Admoni, H. Examining the effects of anticipatory robot assistance on human decision making. In Proceedings of the International Conference on Social Robotics, (ICSR), Golden, CO, USA, 14–18 November 2020; pp. 590–603. [CrossRef]
51. Nikolaidis, S.; Kwon, M.; Forlizzi, J.; Srinivasa, S. Planning with verbal communication for human-robot collaboration. *ACM Trans. Hum. Robot Interact. (THRI)* **2018**, *7*, 1–21. [CrossRef]
52. Ali, M.R.; Van Orden, K.; Parkhurst, K.; Liu, S.; Nguyen, V.D.; Duberstein, P.; Hoque, M.E. Aging and engaging: A social conversational skills training program for older adults. In Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI), Tokyo, Japan, 7–11 March 2018; pp. 55–66. [CrossRef]
53. Saponaro, G.; Salvi, G.; Bernardino, A. Robot anticipation of human intentions through continuous gesture recognition. In Proceedings of the International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, USA, 20–24 May 2013; pp. 218–225. [CrossRef]
54. Ondáš, S.; Juhár, J.; Kiktová, E.; Zimmermann, J. Anticipation in speech-based human-machine interfaces. In Proceedings of the 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Budapest, Hungary, 22–24 August 2018; pp. 000117–000122. [CrossRef]
55. Tellex, S.; Gopalan, N.; Kress-Gazit, H.; Matuszek, C. Robots that use language. *Annu. Rev. Control Robot. Auton. Syst.* **2020**, *3*, 25–55. [CrossRef]

56. Schydlo, P.; Rakovic, M.; Jamone, L.; Santos-Victor, J. Anticipation in human-robot cooperation: A recurrent neural network approach for multiple action sequences prediction. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; pp. 5909–5914. [CrossRef]
57. Bennett, C.C.; Bae, Y.H.; Yoon, J.H.; Chae, Y.; Yoon, E.; Lee, S.; Ryu, U.; Kim, S.Y.; Weiss, B. Effects of cross-cultural language differences on social cognition during human-agent interaction in cooperative game environments. *Comput. Speech Lang.* **2023**, *81*, 101521. [CrossRef]
58. Bennett, C.C.; Weiss, B.; Suh, J.; Yoon, E.; Jeong, J.; Chae, Y. Exploring data-driven components of socially intelligent AI through cooperative game paradigms. *Multimodal Technol. Interact.* **2022**, *6*, 16. [CrossRef]
59. Hutto, C.; Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proc. Int. AAAI Conf. Web Soc. Media* **2014**, *8*, 216–225. [CrossRef]
60. Park, H.M.; Kim, C.H.; Kim, J.H. Generating a Korean sentiment lexicon through sentiment score propagation. *KIPS Trans. Softw. Data Eng.* **2020**, *9*, 53–60. [CrossRef]
61. Bennett, C.C.; Kim, S.Y.; Weiss, B.; Bae, Y.H.; Yoon, J.H.; Chae, Y.; Yoon, E.; Ryu, U.; Cho, H.; Shin, Y. Cognitive shifts in bilingual speakers affect speech interactions with artificial agents. *Int. J. Hum. Comput. Interact.* **2024**, *40*, 7100–7111. [CrossRef]
62. Bartneck, C.; Kulić, D.; Croft, E.; Zoghbi, S. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* **2009**, *1*, 71–81. [CrossRef]
63. Biocca, F.; Harms, C.; Gregg, J. The networked minds measure of social presence: Pilot test of the factor structure and concurrent validity. In Proceedings of the 4th Annual International Workshop on Presence, Philadelphia, PA, USA, 21–23 May 2001; pp. 1–9.
64. Oh, C.S.; Bailenson, J.N.; Welch, G.F. A systematic review of social presence: Definition, antecedents, and implications. *Front. Robot. AI* **2018**, *5*, 114. [CrossRef] [PubMed]
65. Yum, J.O. The impact of Confucianism on interpersonal relationships and communication patterns in East Asia. *Commun. Monogr.* **1988**, *55*, 374–388. [CrossRef]
66. Ornelas, M.L.; Smith, G.B.; Mansouri, M. Redefining culture in cultural robotics. *AI Soc.* **2023**, *38*, 777–788. [CrossRef]
67. Sabanovic, S.; Bennett, C.C.; Lee, H.R. Towards culturally robust robots: A critical social perspective on robotics and culture. In Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction Workshop on Culture-Aware Robotics (CARs), Bielefeld, Germany, 3–6 March 2014.
68. Clark, L.; Pantidi, N.; Cooney, O.; Doyle, P.; Garaialde, D.; Edwards, J.; Spillane, B.; Gilmartin, E.; Murad, C.; Munteanu, C.; et al. What makes a good conversation? Challenges in designing truly conversational agents. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI), Glasgow, UK, 4–9 May 2019; pp. 1–12. [CrossRef]
69. Edquist, H.; Goodridge, P.; Haskel, J. The Internet of Things and economic growth in a panel of countries. *Econ. Innov. New Technol.* **2021**, *30*, 262–283. [CrossRef]
70. Stanojevic, C.; Bennett, C.C.; Sabanovic, S.; Collins, S.; Baugus, K.; Henkel, Z.; Piatt, J.A. Conceptualizing socially-assistive robots as a digital therapeutic tool in healthcare. *Front. Digit. Health* **2023**, *5*, 1208350. [CrossRef]
71. Xu, W.; Dainoff, M.J.; Ge, L.; Gao, Z. Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI. *Int. J. Hum.–Comput. Interact.* **2023**, *39*, 494–518. [CrossRef]
72. Doncieux, S.; Chatila, R.; Straube, S.; Kirchner, F. Human-centered AI and robotics. *AI Perspect.* **2022**, *4*, 1. [CrossRef]
73. Leite, I.; Pereira, A.; Mascarenhas, S.; Martinho, C.; Prada, R.; Paiva, A. The influence of empathy in human–robot relations. *Int. J. Hum. Comput. Stud.* **2013**, *71*, 250–260. [CrossRef]
74. Häring, M.; Kuchenbrandt, D.; André, E. Would you like to play with me? How robots' group membership and task features influence human-robot interaction. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI), Bielefeld, Germany, 3–6 March 2014; pp. 9–16. [CrossRef]
75. Correia, F.; Alves-Oliveira, P.; Maia, N.; Ribeiro, T.; Petisca, S.; Melo, F.S.; Paiva, A. Just follow the suit! Trust in human-robot interactions during card game playing. In Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), New York, NY, USA, 26–31 August 2016; pp. 507–512. [CrossRef]
76. Rato, D.; Correia, F.; Pereira, A.; Prada, R. Robots in games. *Int. J. Soc. Robot.* **2023**, *15*, 37–57. [CrossRef]
77. Sánchez, P.C.; Bennett, C.C. Facial expression recognition via transfer learning in cooperative game paradigms for enhanced social AI. *J. Multimodal User Interfaces* **2023**, *17*, 187–201. [CrossRef]
78. Ulmer, J.; Braun, S.; Cheng, C.T.; Dowey, S.; Wollert, J. Human-centered gamification framework for manufacturing systems. *Procedia CIRP* **2020**, *93*, 670–675. [CrossRef]
79. Oliveira, R.; Arriaga, P.; Santos, F.P.; Mascarenhas, S.; Paiva, A. Towards prosocial design: A scoping review of the use of robots and virtual agents to trigger prosocial behaviour. *Comput. Hum. Behav.* **2021**, *114*, 106547. [CrossRef]
80. Zamfirescu-Pereira, J.D.; Sirkin, D.; Goedicke, D.; Ray, L.C.; Friedman, N.; Mandel, I.; Martelaro, N.; Ju, W. Fake it to make it: Exploratory prototyping in HRI. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI), Boulder, CO, USA, 8–11 March 2021; pp. 19–28. [CrossRef]

81. McMillan, D.; Jaber, R.; Cowan, B.R.; Fischer, J.E.; Irfan, B.; Cumbal, R.; Zargham, N.; Lee, M. Human-Robot Conversational Interaction (HRCI). In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI), Stockholm Sweden, 13–16 March 2023; pp. 923–925. [CrossRef]
82. Ruijten, P.A.; Midden, C.J.; Ham, J. Ambiguous agents: The influence of consistency of an artificial agent's social cues on emotion recognition, recall, and persuasiveness. *Int. J. Hum. Comput. Interact.* **2016**, *32*, 734–744. [CrossRef]
83. Ghazali, A.S.; Ham, J.; Barakova, E.; Markopoulos, P. The influence of social cues in persuasive social robots on psychological reactance and compliance. *Comput. Hum. Behav.* **2018**, *87*, 58–65. [CrossRef]
84. Bartneck, C.; Soucy, M.; Fleuret, K.; Sandoval, E.B. The robot engine -Making the unity 3D game engine work for HRI. In Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Kobe, Japan, 31 August–4 September 2015; pp. 431–437. [CrossRef]
85. Villa, S.; Mayer, S. Cobity: A Plug-And-Play Toolbox to Deliver Haptics in Virtual Reality. In Proceedings of the Mensch und Computer, Darmstadt, Germany, 4–7 September 2022; pp. 78–84. [CrossRef]
86. Kraft, M.; Rickert, M. How to teach your robot in 5 minutes: Applying UX paradigms to human-robot-interaction. In Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Lisbon, Portugal, 28 August–1 September 2017; pp. 942–949. [CrossRef]
87. Butters, D.; Jonasson, E.T.; Pawar, V.M. Exploring effects of information filtering with a VR interface for multi-robot supervision. *Front. Robot. AI* **2021**, *8*, 692180. [CrossRef]
88. Walker, M.; Phung, T.; Chakraborti, T.; Williams, T.; Szafir, D. Virtual, augmented, and mixed reality for human-robot interaction: A survey and virtual design element taxonomy. *ACM Trans. Hum. Robot. Interact. (THRI)* **2023**, *12*, 1–39. [CrossRef]
89. Prattico, F.G.; Lamberti, F. Mixed-reality robotic games: Design guidelines for effective entertainment with consumer robots. *IEEE Consum. Electron. Mag.* **2020**, *10*, 6–16. [CrossRef]
90. Groechel, T.; Shi, Z.; Pakkar, R.; Matarić, M.J. Using socially expressive mixed reality arms for enhancing low-expressivity robots. In Proceedings of the IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), New Delhi, India, 14–18 October 2019; pp. 1–8. [CrossRef]
91. Holz, T.; Dragone, M.; O'Hare, G.M. Where robots and virtual agents meet. *Int. J. Soc. Robot.* **2009**, *1*, 83–93. [CrossRef]
92. Al Moubayed, S.; Beskow, J.; Skantze, G.; Granström, B. Furhat: A back-projected human-like robot head for multiparty human-machine interaction. *Cogn. Behav. Syst.* **2012**, 114–130. [CrossRef]
93. Agarwal, P.; Al Moubayed, S.; Alspach, A.; Kim, J.; Carter, E.J.; Lehman, J.F.; Yamane, K. Imitating human movement with teleoperated robotic head. In Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), New York, NY, USA, 26–31 August 2016; pp. 630–637. [CrossRef]
94. Blasi, D.E.; Henrich, J.; Adamou, E.; Kemmerer, D.; Majid, A. Over-reliance on English hinders cognitive science. *Trends Cogn. Sci.* **2022**, *26*, 1153–1170. [CrossRef] [PubMed]
95. Bennett, C.C.; Sabanovic, S. Deriving minimal features for human-like facial expressions in robotic faces. *Int. J. Soc. Robot.* **2014**, *6*, 367–381. [CrossRef]
96. Deng, E.; Mutlu, B.; Mataric, M.J. Embodiment in socially interactive robots. *Found. Trend Robot.* **2019**, *7*, 251–356. [CrossRef]
97. Choi, H.; Crump, C.; Duriez, C.; Elmquist, A.; Hager, G.; Han, D.; Hearl, F.; Hodgins, J.; Jain, A.; Leve, F.; et al. On the use of simulation in robotics: Opportunities, challenges, and suggestions for moving forward. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e1907856118. [CrossRef] [PubMed]
98. Ter Stal, S.; Broekhuis, M.; van Velsen, L.; Hermens, H.; Tabak, M. Embodied conversational agent appearance for health assessment of older adults: Explorative study. *JMIR Hum. Factors* **2020**, *7*, e19987. [CrossRef]
99. Loveys, K.; Sebaratnam, G.; Sagar, M.; Broadbent, E. The effect of design features on relationship quality with embodied conversational agents: A systematic review. *Int. J. Soc. Robot.* **2020**, *12*, 1293–1312. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.