Multimodal Technologies and Interaction

MDPI

Article

Exploring Data-Driven Components of Socially Intelligent AI through Cooperative Game Paradigms

Casey Bennett [1,2,*], Benjamin Weiss [3], Jaeyoung Suh [1], Eunseo Yoon [1], Jihong Jeong [1] and Yejin Chae [1]

[1] Department of Data Science, Hanyang University, Seoul 04763, Korea; donddog@hanyang.ac.kr (J.S.); hanid99@hanyang.ac.kr (E.Y.); jjh103@hanyang.ac.kr (J.J.); truth01@hanyang.ac.kr (Y.C.)
[2] College of Computing and Digital Media, DePaul University, Chicago 60601, IL, USA
[3] Quality and Usability Lab, Technische Universität, 10623 Berlin, Germany; benjamin.weiss@tu-berlin.de
* Correspondence: cabennet@hanyang.ac.kr

Abstract: The development of new approaches for creating more "life-like" artificial intelligence (AI) capable of natural social interaction is of interest to a number of scientific fields, from virtual reality to human–robot interaction to natural language speech systems. Yet how such "Social AI" agents might be manifested remains an open question. Previous research has shown that both behavioral factors related to the artificial agent itself as well as contextual factors beyond the agent (i.e., interaction context) play a critical role in how people perceive interactions with interactive technology. As such, there is a need for customizable agents and customizable environments that allow us to explore both sides in a simultaneous manner. To that end, we describe here the development of a cooperative game environment and Social AI using a *data-driven approach*, which allows us to simultaneously manipulate different components of the social interaction (both behavioral and contextual). We conducted multiple human–human and human–AI interaction experiments to better understand the components necessary for creation of a Social AI virtual avatar capable of autonomously speaking and interacting with humans in multiple languages during cooperative gameplay (in this case, a social survival video game) in context-relevant ways.

Keywords: human–robot interaction; social cognition; cooperative games; speech systems; virtual avatar; autonomous agents

Citation: Bennett, 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. https://doi.org/10.3390/mti6020016

Academic Editor: Roger K. Moore

Received: 21 January 2022
Accepted: 15 February 2022
Published: 17 February 2022

1. Introduction

There are many ways in which two agents (whether artificial or human) can interact, from cooperatively to competitively and many variations in between. At the same time, there is a growing interest in understanding how we can create artificial intelligence (AI) that emulates natural human social behavior in diverse settings to produce better interactive technology [1]. Those various types of interactions provide different arenas to explore that question, such as through competitive or cooperative game play. One advantage of cooperative game paradigms is that they expose the constraints between the designer's conception of sociality in AI and the user's embedded expectations, since in a cooperative game environment, the AI agent and human user must work together to achieve some goal in a "coordinated" fashion [2]. In particular, those constraints include indirect non-verbal aspects of the otherwise direct verbal interactions that form a core part of the interaction context within which the communication content must be understood. Indeed, a misalignment between context and content is often the basis for communication breakdown [3].

Many of us have had this experience when speaking a second language which is not our mother tongue. There are a host of subtle cues that create *social fluidity* which go beyond simply making the appropriate response in terms of verbal content, such as timing, cadence, appropriate social pauses, and spatial behavior during the communicative

context [4,5]. In addition, non-verbal behavior can be considered as a communicative signal itself, and should be regarded as such in conjunction with verbal actions, since verbal and non-verbal action may be expected to co-occur and at times replace each other [6,7]. There also exists a need for a delicate balance between predictability and unpredictability as it relates to the mental model of one interaction partner toward the other [8]. As shown via the Turing-test-based Loebner Prize competition, an agent that does the same thing over and over is not perceived as "alive" versus an agent that behaves unpredictably at times [9,10]. However, we know from social psychology as well as work on social robots that a certain amount of predictable structure is necessary in social interactions, particularly when there are failures or miscommunication [11].

Previous work has focused on understanding how such fluidity arises from the construct of *social presence*, a sense of being there with a "real person" in artificial environments [12]. Such previous research has shown that both behavioral factors related to the artificial agent itself as well as contextual factors beyond the agent (i.e., interaction context) play a critical role in how people perceive interactions with interactive technology [13]. Cooperative game paradigms offer us an opportunity to expand upon this by manipulating the interaction context in specific ways, particularly if we develop a "Social AI" with components that we can modify on demand. By simultaneously manipulating both the game environment and the Social AI, we can explore a fuller range of questions around which components of social interaction—both behavioral and contextual—are relevant to producing the social fluidity necessary for humans to perceive an interaction as "life like". These questions also tie back to Dennett's work on intentional stance as it relates to attributions of agency in artificial agents, i.e., an agent that is perceived to have its own self-driven goals and intentions (averse to a machine) [14].

In order to do this effectively, however, requires us to develop methods for designing the AI agent and cooperative game play environment based on relevant data about interactions within said environment. Here, we describe an approach for *data-driven development* of a Social AI and cooperative game environment along those lines, utilizing both human–human interaction data and human–AI interaction data. In this case, the Social AI took the form of a virtual avatar capable of autonomous speech based on its perceptions of the social environment, which was a customizable social survival video game.

This paper is organized as follows. Section 2 describes prior related work in the existing scientific literature. Section 3 describes the methods, cooperative game paradigm, and experimental design for the current research. Section 4 details the results of those experiments. Section 5 discusses the implications of those results and potential future research avenues.

2. Prior Work

2.1. Social AI in Human–Robot Interaction

For the case of embodied agents, the interaction context is typically a physical environment that generates multi-modal data, including motion, sound, and touch. One useful source of evidence for how agents should interact socially in this multi-modal sense can be drawn from the existing literature on human–robot interaction (HRI) [15]. For instance, evidence from robotic faces has shown that different interaction components, such as visual appearance and sound, can have a significant impact on human perceptions of the interaction [16]. External context *outside* the face itself (such as simultaneously occurring movie clips) can alter the perception of the robot's facial expression to the point that the human "sees" completely different expressions due to the context even when the expression itself is the exact same [17]. Moreover, users have specific preferences for the capabilities a social robot should have [18], both in terms of the robot's behavior and responses to human actions in certain situations [19].

Other research has shown similar context effects during studies of robots engaged in physical game playing [20,21], as well as effects from robot group size [22]. The contextual factors in those cases included not only the physical embodied form of the robots, but

also their behaviors and ability to elicit trust/empathy from human interactors in order to accomplish shared goals. In these embodied paradigms, we can see the impact of many different types of interaction components, although the ability to rapidly alter the physical agent and/or physical environment is a limiting factor in exploring variations of those components in embodied settings. Hence, there is a need for other research paradigms to expand upon such findings.

2.2. Interactive Speech Systems

There is also significant research relevant to Social AI that can be drawn from interactive speech systems, which seek to interactively engage users conversationally in order to accomplish some task or goal. Task-oriented speech systems thus have, by definition, a topic or domain to converse about. Therefore, designers try to consider most of the technically available information within that domain in order to mimic human speech, such as location, dialog history, user identity, and other contextual information [23]. On the one hand, this aim is to facilitate users resolving underspecified references that may have multiple meanings depending on context ('it', 'there', 'the second', 'tomorrow'). On the other hand, a secondary aim is to enable the design of agents that can utilize shorter, more natural dialog. This comprises the reverse, where the agent infers information not explicitly stated by the user, such as the current status of the world, situational facts, or user preferences, in order to establish a dialogue state within which it can appropriately act and fluidly respond, e.g., recovering from repeated conversation errors and/or adapting to linguistic conformations. Such established design aspects affect user experience in terms of perceived cooperativity and appropriateness of the conversational agent [24,25].

However, the interplay of an agent's spatial actions and its speech in 3D cooperative environments has not thus far been a big topic in interactive speech systems, apart from some work on multi-modal systems, for which it has yet to be determined whether certain confirmations ('the light [it] is switched on', where the word light could be omitted) or explicit references need to be produced by voice at all rather than non-verbal cues [26]. What is known is that both linguistic (e.g., word choice) and para-linguistic aspects of an agent's behavior may support or hinder the interactive flow, success, and even rapport [27–29]. While non-natural behavior in early task-driven interactive systems mainly affected its user-attributed quality from a physical artifact standpoint [30], modern systems may cause the attribution of intelligence by users so that divergent behavior might be interpreted differently as in during human–human interaction [31]. For example, in humans, a long delay before a positive response can result in the impression of being doubtful [4]. One such established cooperative game-like scenario for spoken interaction is the MAP task, which can identify relations between interaction fluidity and verbal and non-verbal behavior. In this scenario, two players need to solve a navigation task by sharing complementary information [32]. However, since speech-only and, more so, text-only interaction is rather limited in terms of social signals and coordinated actions, speech-based virtual agents hold promise for expanding the study of social fluidity and attributed intentionality.

2.3. Virtual Agents and Virtual Reality

As mentioned in Section 1, the concept of social presence is an important one when dealing with virtual agents, in the sense of "being there" with another intelligent entity [1]. Much prior research has explored this notion by looking at interactions in virtual reality environments. For instance, Slater (2009) has attempted to disentangle what constitutes social presence by separating a human's sense of being in a particular place (place illusion) from the sense that the events are actually occurring regardless of one's own actions (plausibility illusion) [33]. These represent different levels of *immersion* in the interactive experience, with plausibility illusion generally being a higher threshold to achieve. Others have attempted to further explain these constructs by examining the dividing line between what humans perceive as illusion versus reality [34]. Recent research has also explored how

such virtual agents can take on a "social identity" during human–agent interaction, which appears capable of triggering users' self-conceptualization of group social expectations [35].

Critically, we note that the perception of these aspects of social presence with virtual and physical agents has not been found to correlate with physical realism [36], but rather to correlate with perceived social rewards due to the interaction via specific neural pathways [37], or even anxiety about failing to obtain such rewards [38]. To put it more simply, there is an emotional aspect to this, which can often override other cognitive processes. Interestingly, recent research has found that given appropriate contextual factors and agent behaviors, social presence in virtual reality environments can be the same in completely implausible environments as in more realistically plausible ones [39], further indicating the importance of considering the design of both the agent itself as well the interaction context.

2.4. Mental Models of AI and Game Theory

There are other lines of research that can contribute to our understanding of contextual factors on human perceptions of interactions with AI agents. For instance, Gero (2020) explored the mental models human users develop during interactions with AI during cooperative word games. Their findings suggest that understanding these mental models is key to developing more self-explanatory AI systems, rather than just having some opaque reinforcement learning or deep learning model controlling the AI [2]. Similarly the emerging field of explainable AI seeks to develop methods to provide cognitive scaffolding that more closely aligns human thought processes with the inner workings of AI models [40]. More broadly, game theory in the field of economics has long sought to explain the emergence of optimal behavior given different contexts, including the differences between cooperative and competitive scenarios [41]. In short, that research underscored how different contexts demand different agent behavior, as well as how the interplay between the two (agent behavior and interaction context) plays a critical role in human perceptions of the situation [42,43].

The challenge remains to integrate these seemingly disparate lines of research described here in Section 2 into flexible paradigms to explore different possible explanations for social interaction with AI, in terms of both the AI agent and the interaction context. Such research paradigms can supplement the above approaches, as well as potentially address some of their limitations. Hence, that need is the motivation for the work described in this paper.

3. Methods

3.1. Cooperative Game Environment

In the current work, we utilized a video game called *Don't Starve Together* for our cooperative game environment (https://www.klei.com/games/dont-starve-together, accessed on 20 January 2022) which can be downloaded from online sources such as Steam. The *Don't Starve Together* game is a social survival game where players need to collect resources, make tools, fight monsters, and cooperate with each other to survive longer (visual examples can be seen in Section 3.2 below). Much like other "crafting games" such as the popular Minecraft game, *Don't Starve* requires players to collect specific combinations of resources in order to build things, without which they will be vulnerable to various dangers and likely lose the game via player death. The pressure to accomplish those tasks is under time constraints, as the level of dangers gradually increases over time. The game is heavily customizable through the use of game modification tools (henceforth referred to as the "Game Mod"), which allow users to alter the mechanics of the in-game environment and non-player character (NPC) behaviors through the LUA programming language. Additionally game mods can be shared by users through an online, open-source community. As such, the game provides an ideal environment to experiment with interactive behavior during cooperative goal-oriented tasks [44,45].

The game can function in single-player or multi-player mode (up to 6 players simultaneously), though there were always only 2 players (the avatar player and human player) in

the experiments described here. Experiment game sessions were conducted via a secure "friends only" room on the *Don't Starve* online servers to prevent any interruption in the test, allowing for uninterrupted 30 min one-on-one gameplay sessions. Sessions were conducted from two separate computers in separate locations both running the game through the same server, henceforth referred to as the "confederate" computer and the "participant" computer. The avatar player was controlled from the confederate computer, while the participant computer was used for the human player. This is described in more detail in Section 3.2.

For our experiments, a custom Game Mod was developed with two purposes in mind. First, we wanted to create "game data writing" functionality, so that we could collect real-time data about the game state at every moment. This included information about player status, inventory, movement, items equipped, attacking/fighting, time of day, and entities in the players' immediate environment (e.g., monsters, structures). Such data was fundamental to our various analysis of social interactions during the game and the results described in Section 4. Second, we wanted to create custom game environments in order to be able to conduct controlled experiments. For instance, this included creating a fixed starting position with various resources immediately available ("advanced starting position"), providing a constant source of light at that starting position in order to encourage players to return to it periodically to encourage more social interaction ("base camp"), and setting the minimum health level at 10% so that players could not die, guaranteeing every experiment game session could last approximately the same amount of time, e.g., 30 min ("partial invincibility"). However, for our experimental design, human players were given some general directions on how gameplay works, but not informed of these Game Mod features.

We did make additional use of a few community game mods. This included mods allowing us to pause the game to control the start of the session, as well as the use of the FATIMA toolkit to experiment with AI development so that the game can be played autonomously by an AI agent in the future. We describe this latter AI component in more detail in Section 4.1.

3.2. Experiments

In order to collect data for our data-driven development of the Social AI, we conducted a series of focused experiments to progressively test components. This included a first experiment (n = 6) and a second experiment (n = 8), comprising 6 males and 2 females (only 4 of the males participated in the first experiment). A total of 14 game sessions were conducted with two players at a time, each approximately 30 min in duration. Additionally, these participants were split into 4 from Western countries (from the US and Germany) and 4 from East Asian countries (from Korea), as part of our long-term goal here is to conduct cross-cultural comparisons of Social AI interactions. Likewise, the speech systems developed for the Social AI are multi-lingual, capable of understanding and speaking in both English and Korean languages. These were part of a series of experiments (approved by the Hanyang University IRB, #HYU-2021-138).

For the first experiment, the goal was to collect data about naturalistic human vs. human gameplay. In this scenario, we setup a Zoom meeting to allow direct audio–visual communication between the humans while playing the game, in a side-by-side configuration. We then used OBS Studio (https://obsproject.com/, accessed on 20 January 2022) to record the entire screen during gameplay, including the game window itself as well as the Zoom window of simultaneous social interactions. An example of this can be seen in Figure 1. For the second experiment, the human on the "confederate" computer side (see Section 3.1) was replaced by a virtual avatar. For this, we linked the written game data from the Game Mod to a Social AI, capable of reacting to in-game events through autonomously generated speech. This Social AI was based on data collected in the first experiment, written in the Python programming language. We used locally-installed (Window or Mac) voice packages as part of the Text-to-Speech (TTS) module, with the

audio output redirected to an internal "virtual" microphone jack, and then used the Loomie application (https://www.loomielive.com/, accessed on 20 January 2022) as a visual avatar capable of moving its lips synchronously with the speech. The second experiment was then conducted using a Wizard-of-Oz (WoZ) design, where the virtual avatar socially interacted in an autonomous manner with the human player during gameplay, but its in-game actions were still controlled by a human from the confederate computer. In the future, we plan on replacing this WoZ setup with an AI agent to control the in-game actions as well (see Discussion section). Otherwise, the experiment was the same as the first, with the avatar player and human player interacting through a Zoom window, and the entire screen being recorded via OBS. An example of this can be seen in Figure 2. We also include a brief 1 min video of human–agent *autonomous* speech interactions between the AI agent and participants during gameplay in Supplementary Video S1, with several simple examples.

Figure 1. Gameplay example during the first experiment (human vs. human).

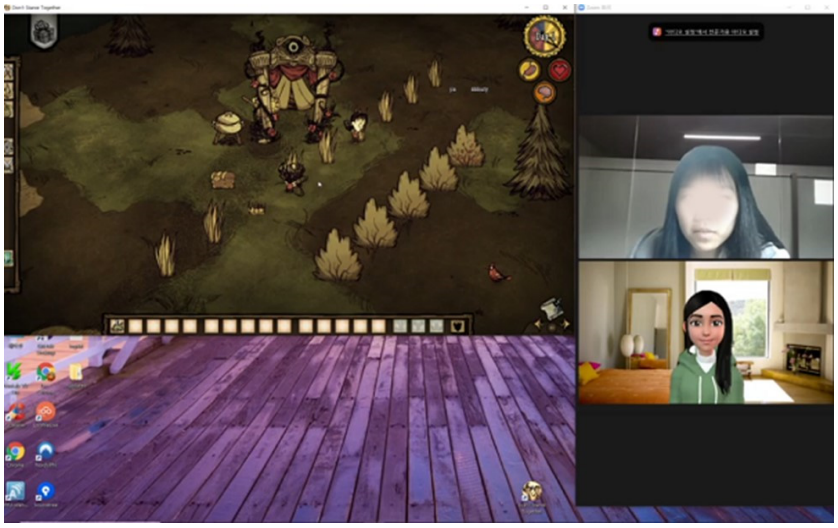

Figure 2. Gameplay example during the second experiment (human vs. avatar).

During each experiment, several types of data were collected. This included audio-video recordings from OBS studio, which were used to create speech annotations and for facial/gestural recognition analysis. We also collected the written game data, which was generated approximately twice per second and timestamped so that it could be matched to

the audio-video recordings for further analysis. This allowed us to connect in-game events to social interactions occurring around the same time frame. Finally, we also collected qualitative data about the participant experience using a questionnaire (available from the main author's website), which included questions about when they felt like the Social AI speech matched the gameplay (or did not), what they found annoying, and issues with the Game Mod and/or social environment. This allowed us to identify components that needed to be augmented in the future, from the participant's perspective.

4. Results

4.1. Speech Hierarchy Development

One critical component for interaction with a Social AI is development of speech systems, so that an autonomous agent can interact verbally with human participants [46]. In order to accomplish this, it is necessary to create *context-specific* content geared toward the task (as mentioned in Section 2), which in our case is a cooperative game scenario. This was accomplished by generating speech annotations from the first experiment (human–human), which were then augmented after the second experiment (human–avatar). The process was as follows.

First, the audio–visual recordings were analyzed to produce a list of common player utterances (occurring more than once), linked to the game situation in which the communication occurred. The situations comprised various aspects related to the game, such as game events changing the status of resources, monsters in the vicinity, and activities such as making tools. Based on these annotations, a hierarchy diagram was derived to specify each situation and co-occurring speech utterances. The hierarchy was derived by four separate human coders, who first categorized the utterances independently for both English-language and Korean-language videos, and then worked during a focus group to align those categories into a hierarchy. The full hierarchy is too large to present here, but an example of one part of the hierarchy (Monster-related) can be seen in Figure 3, with some example utterances in the leaf nodes.

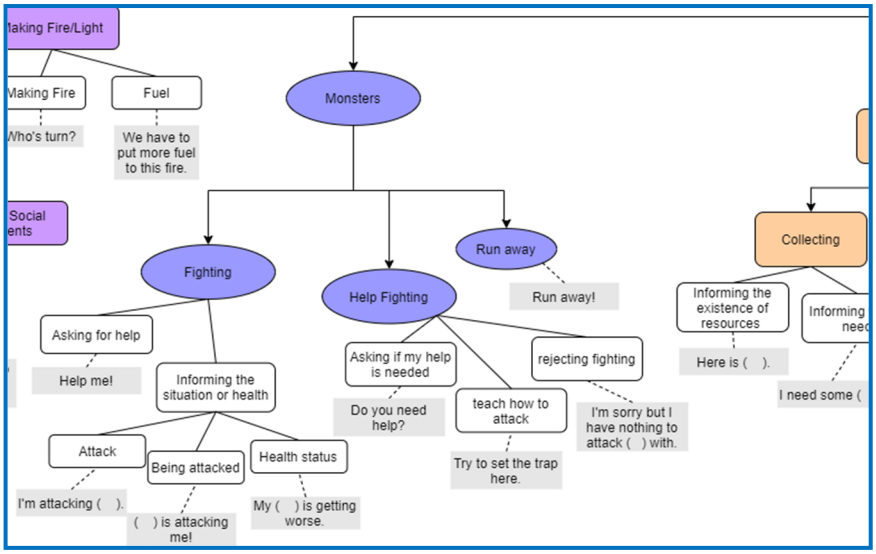

Figure 3. Part of the speech hierarchy (using "Monsters" section as an example).

At a high level, these events were categorized into various game states: Build Stuff, External Events, Monsters, Resources, Social Interaction, and Night. Each of these high-level game states (6 total) were comprised of various sub-categories (18 total), which were then further broken down into various branches. Interrater reliability (Cohen's Kappa) for these video annotations by different coders was calculated as 0.718, which indicates strong agreement among the coders even across the different languages. Given the similarities, the developed hierarchy was identical for both English and Korean speech systems, so

that the content matched on both sides though there were some differences in phrasings of specific utterances (due to linguistic differences). An initial version of this was developed after the first experiment and then tested during the second experiment. The findings from the second experiment led to the results described in Sections 4.2 and 4.3 below.

We also note that this hierarchy was used as the basis for development of an AI agent to control in-game actions for fully autonomous gameplay via the FATIMA toolkit. Each category and/or sub-category in the hierarchy became the basis of various "use cases" that we can use to define capabilities and goals for the AI agent, so that the autonomous social interaction of the virtual avatar and in-game actions are aligned, without the need for a WoZ experimental design. During the experiments described in this paper, this component was not yet implemented, but we return to this topic in the Discussion section.

4.2. Interaction Component Development

During the second experiment, we focused on identification of missing *interaction components* beyond the speech system content itself. In particular, the goal was to generate data for designing these components in a data-driven manner, rather than based on pre-conceived notions from a "designer's perspective". These components were first identified by analyzing interaction questionnaires collected during the experiments, from which we extracted lists of participants' frequent statements about missing and/or important "components" during interaction with the Social AI. These items include:

- Responsiveness to human player communication (including direct questions);
- More natural sentence variation;
- AI awareness of own recent speech (e.g., not repeating itself too often);
- AI commentary about direct player interactions (e.g., sharing food), rather than just game environment;
- More suggestive speech from the AI (e.g., talking about future plans).

After identifying these items, we utilized the collected gameplay video recordings and written game data to generate a number of interaction components to address them. First, we implemented an automatic speech recognition (ASR) system to create more responsiveness, using the Microsoft Azure speech-to-text API. The ASR was setup to recognize common keywords and phrases spoken by humans, which were extracted from the list of player utterances from the gameplay videos. This included single keywords (such as "monsters" or "food"), as well as combinations of keywords co-occurring anywhere within the same utterance (e.g., "where" and "go", or "monster" and "near"). We then generated a list of 3–5 responses the Social AI could make for each of those keywords. The Social AI could then randomly choose from those responses, if the ASR was triggered. For instance, in the case of the "where/go" keyword combination, the Social AI might say "Lead on, I'll follow you" or "Let's just wander around". This system was implemented so as to match on the English and Korean side, in order to facilitate future cross-linguistic studies.

Along with those ASR responses, there was a need to create more sentence variation in the Social AI's self-generated speech, when it is talking about the game environment on its own rather than responding to a human player. This followed a similar process as the ASR above, but was based on the speech hierarchy utterances (see Section 4.1) rather than keyword responses. Combining both the self-generated utterances from the hierarchy and the ASR response utterances, **this resulted in 46 different utterance categories, with a total of approximately 160 different Social AI responses** to those categories based on the annotated gameplay videos.

In order to create more "speech awareness" in the Social AI and reduce repetitiveness, we adopted an approach utilizing *Social Inhibition of Return* (social IOR), which is based on IOR models from various human sensory functions such as vision [47]. The basic idea here is that there are mechanisms in the brains of naturally intelligent organisms (including humans) that inhibit us from repeating the same behavior in a short period of time (e.g., 2–3 s) in order to maximize task efficiency (e.g., during visual "information foraging") [48]. A failure in these mechanisms is thought to play a role in human mental illness, such as

obsessive-compulsive disorder. In the context of social IOR, these mechanisms are also important to produce fluid natural behavior, rather than repetitive "robot-like behavior" [47]. In order to implement this in our case, we utilized the top-level utterance categories from the speech hierarchy so that the Social AI maintained an internal array to keep track of recently spoken categories, with a "counter" that counted down a certain number of seconds during which any further utterances within that same category were suppressed (though the AI could still make utterances from other categories). Initially, this counter was set to 3 s, based on prior research on social IOR in humans, though whether it should be for shorter/longer time lengths in Social AI needs further research.

Another interaction component derived from the experiment data was "priority levels" for the utterance categories, in order to create a mechanism for controlling the chattiness of the Social AI. The most common participant statement on the questionnaires after the second experiment (8 out of 8, i.e., 100% of participants) was that the avatar sometimes spoke too frequently, even making statements about topics/events not relevant to the task at hand. The priority levels were thus aimed at providing a "dial" to control this, which could be turned up or down like a volume control dial. To do this, we had 6 independent coders who were familiar with the *Don't Starve* video game evaluate a spreadsheet of the utterances and the AI responses to them, assigning each utterance a priority level of either 1 (high), 2 (medium), or 3 (low). These levels were defined for the coders, so:

1. Priority 1:
    a. Direct human questions: any ASR response to those, and
    b. Any "critical" speech that cannot be omitted based on social norms (not replying to "hello").
2. Priority 2:
    a. Fight-related content (attacking, defending), and
    b. Existence-related content (e.g., dying and starving).
3. Priority 3:
    a. Any non-answer ASR response to human speech (i.e., comment not a direct question),
    b. Situational content not related to fighting or existence, and
    c. Anything else that does not fit in priority 1 or 2.

The interrater reliability across coders for when they assigned the same level to each utterance category was calculated as 0.54 (pairwise average Cohen's Kappa) and 0.30 (Fleiss' Kappa for multiple raters), indicating moderate agreement. Meanwhile, Cronbach's Alpha was calculated as 0.789, indicating strong internal reliability for the coding scheme. After coding, we then calculated the average and median of those values for each utterance category, with the resulting values forming the basis of our priority control system by establishing a threshold. Any utterances below that threshold are spoken, while those above are not. Hence, reducing the threshold reduces the chattiness of the Social AI, and vice versa. An example of this for a few of the utterance categories can be seen in Table 1, with each category having 3–5 different responses (not shown for brevity). It remains to be seen via future research whether the average or median produces better performance from the user perspective, but at the time of writing we are utilizing the median. This resulted in 14 utterance categories being identified as higher priority, 16 as medium, and 16 a lower priority.

Table 1. Priority Level Coding (examples).

Type	Utterance Category	Average	Median
ASR	"hello.*": [	1.67	1
ASR	".*food.*": [	1.83	1.5
ASR	".* monster.*near.*": [	2.33	2.5
ASR	".*attack.*": [	1.67	2
ASR	".*where.*go.*": [	1.17	1
ASR	".*night.*": [	2.67	3
ASR	".*campfire.*": [	2.50	3
ASR	".*help.*": [	1.17	1
Self-Generated	"inform_morning": [	3.00	3
Self-Generated	"inform_starving": [	1.50	1.5
Self-Generated	"inform_defense": [	1.67	2
Self-Generated	"inform_torch": [	2.83	3
Self-Generated	"inform_near_light": [	2.50	3
Self-Generated	"inform_only_axe": [	2.67	3
Self-Generated	"inform_a_few_monsters": [	2.33	2.5
Self-Generated	"inform_generic_expression": [	2.83	3

4.3. Deep Learning Models for Interaction Planning

As noted in Section 4.2, one of the findings from the second experiment was that participants reported a need for "planned speech". In other words, one thing a flexible Social AI needs to have is the ability to plan for future interactions, rather than just respond to events that have already occurred. To tackle this challenge, we took the written game data from the second experiment and lined it up with gameplay videos using the ELAN video annotation software (https://archive.mpi.nl/tla/elan, accessed on 20 January 2022). We then matched the timestamps between speech events in the videos related to different game states (see Section 4.1) and the corresponding point in time in the written game data, after which we extracted the game data for the 15 frames prior to the event. As the game data was recorded approximately twice per second, the result was a dataset containing the game data for the 7–8 s leading up to each event. We then filtered that data to remove overlapping speech events, so as to only include isolated events clearly related to the six high-level game states. This resulted in a dataset of 156 game state events based on the speech hierarchy, each with 15 rows of feature data, totaling 2340 total rows. This data can be visualized as a multi-dimensional array, as shown in Figure 4. The goal was to predict utterance categories *before* they occurred, using a data-driven approach rather than a rule-based expectation, so as to allow the Social AI to speak about impending events likely to occur in the immediate future (i.e., planned speech).

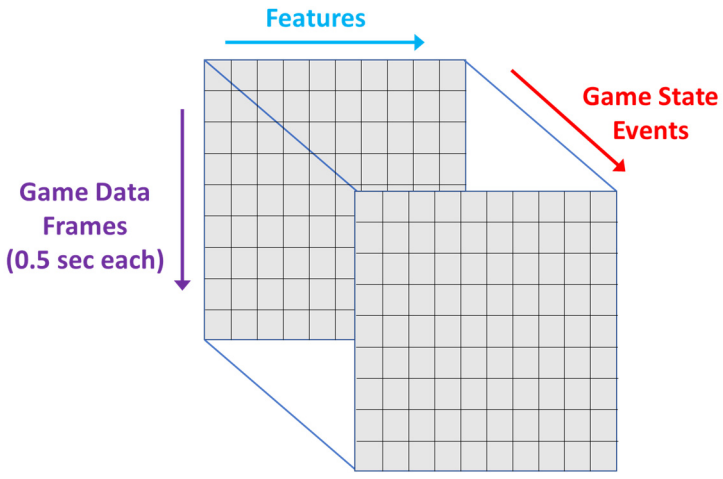

Figure 4. Keras input example for predicting game events.

To model this data, we utilized both deep learning (DL) and standard machine learning (ML) approaches. DL was performed using the Python package Keras (https://keras.io/, accessed on 20 January 2022), which is a deep learning library based on TensorFlow. The game state events became the targets, while the game data was used as the features. The resulting data was then fed into a deep learning model consisting of a single 1D convolutional neural network layer (CNN) with kernel size set to 1 and using a ReLU activation function, followed by a single recurrent layer (LSTM) with 30 units [49]. The idea was that the CNN could parse out "invariant representations" of pattern signatures occurring anywhere in the interaction, followed by the LSTM detecting critical "sequences" of those patterns over time. A final fully connected "Dense" layer using a sigmoidal activation function was used to make the final binary classification predictions. To evaluate performance, 20% of the data was held out as a "test set" for each classification run. Due to target class imbalance, training data was re-balanced using SMOTE [50].

We also attempted standard ML models using the Python package Scikit-Learn (https://scikit-learn.org, accessed on 20 January 2022). Multiple modeling methods were attempted: random forest, gradient boosting, and support vector machines. Models were generally run using the default parameters in Scikit. Performance was estimated using multiple performance metrics (e.g., accuracy and AUC), based on 5-fold cross validation, following standard machine learning guidelines [51,52]. In order to predict the game state target using the standard ML approach, feature data was "collapsed" into aggregated data across each 15 min interaction by calculating averages/percentages for each feature across the entire window, resulting in a single row of data for each game state target.

Results can be seen in Table 2 for both the DL and ML models (RF = random forests, GB = gradient boosting, SVM = support vector machines, and DL = deep learning). In general, predictions worked for some game states, but not others. In particular, it was difficult to predict a social interaction (e.g., random banter) before the social interaction occurred, unless it was tied to some specific game situation (e.g., monsters). Likewise, resource collection was also difficult to predict. The ML models using collapsed data had lower performance than the DL models, though the average differences were small, particularly in terms of AUC. This may indicate that there were no significant temporal sequences in the data leading up to the event that were detectable via recurrent models. However, it is also possible that we did not have the correct data fields in our dataset for that to work properly, given that there are potentially thousands of variables that could be included from the game environment while we focused on only 30 of them here. This is an area that needs further research, but these initial results suggest that planned speech may be possible in such cooperative game environments.

Table 2. Planned Speech Prediction Results.

	RF		**GB**		**SVM**		**DL**	
Type	**Acc**	**AUC**	**Acc**	**AUC**	**Acc**	**AUC**	**Acc**	**AUC**
Night	0.81	0.9753	0.78	0.9182	0.75	0.7697	0.87	0.9224
Resources	0.42	0.5117	0.46	0.5175	0.47	0.4792	0.59	0.6446
Monster	0.79	0.8980	0.75	0.8562	0.69	0.7907	0.85	0.9214
Build Stuff	0.89	0.9561	0.83	0.9539	0.82	0.9621	0.88	0.9033
Social Interaction	0.52	0.5080	0.52	0.4844	0.55	0.6565	0.50	0.5015
External Events	0.86	0.9656	0.81	0.9604	0.81	0.9632	0.84	0.9546
Average	*0.72*	*0.8025*	*0.69*	*0.7818*	*0.68*	*0.7702*	*0.76*	*0.8080*

4.4. Facial and Gestural Recognition

Another aspect of Social AI development is utilization of multi-modal cues, such as facial expression and gestural recognition. To that end, we analyzed the data from the second experiment to evaluate the potential for identifying participant facial expressions while interacting with the avatar during gameplay for the seven basic Ekman facial

expressions: Happy, Sad, Surprising, Angry, Fearful, Disgusting, and Neutral [53]. To do so, we first used the same ELAN annotation software as in Section 4.3 to annotate facial expressions during gameplay videos using a human coder. After that, we used the Python package OpenCV (https://opencv.org/, accessed on 20 January 2022) to detect human faces anywhere in the scene via Haar Cascades. We then utilized a pre-built CNN model for facial expression recognition from Keras (see Section 4.3) and applied that to our videos to produce a probability value of each of those seven facial expressions for a given video frame.

Based on the video annotations, we can see that most of the time users displayed a "neutral" facial expression (Figure 5), which is not surprising since they were focused on the game itself. A small minority of the time they displayed other facial expressions such happy, sad, and surprise. As such, we attempted several types of *criteria* to obtain accurate recognition. This involved calculating a predicted probability for each facial expression for every video frame from the CNN, looking across multiple frames comprising each second (typically about 30 frames per second) in order to smooth performance, and then choosing the most frequently detected expression within that one-second window. However, expressions were only considered as detected if they fit the criteria conditions, e.g., having probability above some threshold. The idea was to minimize false positives, and simply assume a neutral facial expression unless we were reasonably certain otherwise.

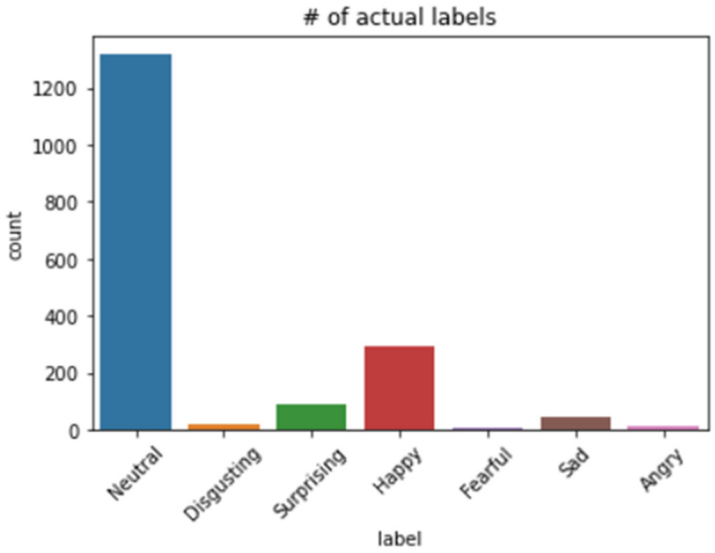

Figure 5. Average distribution of facial expressions per video.

The analysis evaluated a number of different criteria conditions in order to compare performance (Table 3). This included simply choosing the facial expression with the maximum probability, regardless whether that was 30% or 90% (Criterion 1). We also tried only choosing the maximum expression above a certain set probability threshold, otherwise defaulting to predicting the facial expression as neutral if none exceeded the threshold (Criterion 2). We then added a restriction to the first two criteria so that the prediction of the first and last frame of the window had to match (Criteria 3 and 4). For criteria 1–4, we also evaluated weighting the predicted probabilities by the overall average frequency of each expression (as shown in Figure 5), so that expressions that occurred more frequently in general (e.g., happy, surprising) were given more weight when determining the highest probability for a particular window. Finally, we created another criterion that utilized the predictions from all the previous criteria (1–4) as features, and then attempted to predict the facial expression target via a random forest model (Criterion 5). In short, this final criterion attempted to combine different thresholds and weighting schemes into a single prediction, akin to how multiple filters might be used in a CNN deep learning model.

Table 3. Facial expression recognition by criteria (o = weighted, x = unweighted).

Name	Description	Weighted	Threshold	Accuracy
Criterion 1	Choose Max Probability	x	-	9.2
		o	-	29.7
Criterion 2	Choose Max above Threshold	x	60	13.7
		x	70	19.0
		x	80	24.5
		o	60	42.9
		o	70	43.0
		o	80	34.7
Criterion 3	Crit. #1 + First/Last Frame Match	x	-	34.7
		o	-	62.8
Criterion 4	Crit. #2 + First/Last Frame Match	x	60	44.9
		x	70	50.7
		x	80	56.5
		o	60	70.2
		o	70	70.2
		o	80	66.1
Criterion 5	Combined Criteria Pred Model	-	-	80.1

The maximal performance of this approach was 80.1% accuracy using Criterion 5, which is significantly above random chance (14.4%, or 1/7) but perhaps not optimal. Others have reported performance in the mid-80 percent range during video game play [54,55]. There were a number of challenges we observed which potentially contributed to this. In particular, during cooperative game play in these virtual "social survival" games, users are often focusing on survival and so have a look of "concentration" that resembles a neutral expression. Direct face-to-face interactions seem to be sparse and linked to sporadic game events. This seems to be a positive in one sense, in that it allows us to focus on speech and other non-verbal cues, rather than facial expressions. However, at the same time, it can be seen as a weakness, whereas in-person interactions with embodied agents may be different from these virtual interactions in substantive ways [33]. A question remains as to whether this is reflective of virtual technology interactions in general or just our experimental paradigm. It is also possible that developing a customized Keras CNN model for our specific game scenario may work better than a generalized pre-trained model we used for this work, but such development would require the creation of a large *task-specific* corpus for training as is typical with many deep learning models, which would then be limited to only this specific task environment. Along with gestural recognition, this functionality is still being explored and will likely be part of further future work (see Discussion).

5. Discussion

5.1. Main Summary

This paper explored the use of cooperative game paradigms to better understand how we can create more life-like Social AI in time-constrained, task-oriented environments [2]. Cooperative game paradigms offer an ideal environment to explore the interplay of the *interaction context* with agent behavior during human–agent interaction [13], since both are interdependent, particularly when the human and agent must collaborate to achieve some goal under time pressure [1]. To this end, **we focused here on the exploration of methods for developing modifiable components of a Social AI and game environment** to enable the simultaneous manipulation of both the agent and the game environment. In particular, we evaluated how a data-driven approach could be used to develop various components, utilizing both human–human interaction data and human–AI interaction data.

Results showed successes as well as areas of improvement for different components. We were able to develop a multi-lingual *content-specific* speech hierarchy for the Social AI based on human–human gameplay (Section 4.1), representing dozens of high-level

game states and various sub-states, with high reliability (Cohen's Kappa, 0.72). The speech hierarchy became the basis for the agent's autonomous speech system during social interaction (in both English and Korean), as well as "use cases" for an AI controller for autonomous gameplay. **Subsequent human–agent gameplay experiments using this system revealed a number of missing "interaction components"**, such as responsiveness to human player communication and AI self-awareness, which were then addressed via various data analyses (Section 4.2). For instance, analysis of human utterances to the agent during gameplay led to the creation of an ASR system to respond directly to human questions/comments. Other needed components included one based on the principle of social IOR, utilizing the neuroscientific concept of "inhibition of return" in natural intelligence, to limit repetitive social interaction [47], as well as a "priority system" to control the level of interaction (e.g., chattiness of the agent). Analysis of these components showed strong internal consistency and reasonable reliability (Cronbach's Alpha 0.79, Cohen's Kappa 0.54).

Along with the above, additional data analysis was undertaken to create "planned speech" based on machine learning models using game data to predict events likely to occur in the *immediate* future (Section 4.3), so that the Social AI could engage in anticipatory talk about impending events rather than only what has already occurred in the past or current situation. To enhance speech interactions, we also undertook facial expression recognition analysis of human participants during gameplay in order to detect emotions during gameplay (Section 4.4), combining several different threshold and weighting schemes together to achieve 80% accuracy without the need for a task-specific corpus. Meanwhile, we implemented "Game Mod" functionality that allowed us to create a customized social environment and interaction scenarios, in order to conduct controlled experiments around different hypotheses related to these components of Social AI.

Critically, we note that many of the components explored here (e.g., planned speech and social IOR) are **geared toward expanding the multi-modal nature of Social AI interaction, both in a verbal/non-verbal and temporal sense**. These components have not been identified with the traditional aim of optimizing usability or user experience of an interactive spoken device [56], but rather with the goal of fostering intentionality attribution in an autonomous agent. While these aims do not exclude each other, our approach resulted in a specific focus on appropriateness and agency, rather than effectiveness or efficiency. Moreover, many of the components are directly tied to the cooperative nature of the social environment as well, which underscores the interplay of the AI behavior and contextual factors [2]. Indeed, the component capabilities are deeply interlinked to the characters' cooperative actions and game session evolvement—a known pre-requisite for creating a successful interactive conversational agent [57]. This situatedness demands an empirical (in our case data-driven) design approach, which aligns with best practices for designing immersive voice interaction [21]. It also highlights how that same process can be used to create customizable social environments to explore a broad range of hypotheses related to how contextual factors relate to people's perceptions of interactive technology.

5.2. Future Work and Broader Impact

The work here highlights how different components of a social agent (whether virtual or physically embodied) and its interaction context can impact the fluidity of the social interaction and the subsequent sense of *social presence*, i.e., the sense of "being there" with a real person in a given environment [12]. Future work is needed though to understand the specific effects of individual components of both the agent and context, and moreover how they might affect each other. We are currently conducting large-scale trials of hundreds of participants to answer some of those questions, by meticulously turning individual components on and off. However, potential questions in that regard are vast, and will likely go beyond any individual research group's work or interaction modality (speech, facial expressions, non-verbal cues, etc.). Furthermore, there is additional technical work to be carried out, such as replacing the current ASR keyword implementation with more

sophisticated information extraction techniques (e.g., intent and entity recognition) [58]. Beyond the facial expression recognition lies an opportunity to integrate gestural recognition and other types of multi-modal affective interaction which were not addressed here [59]. There is also a need to further develop the AI for autonomous gameplay to control in-game actions alongside the autonomous interaction components to get away from the limitations of WoZ experimental designs [60]. This may allow for a closer exploration of how the agent's in-game actions and its social interaction behaviors might best be aligned with the interaction context, in a triadic sense.

Such questions have a direct bearing on the shape of future interactive technology in society, in particular how we might best utilize the data from those interactions. Often times, such data is seen as a closed-form solution to solve an individual design problem; however, the impacts can go beyond the specific design problem itself. For instance, one particular impact may be *technology accessibility*. Many of the ideas explored in this paper are geared toward creating greater fluidity in the interactions between agents and human users, but the ultimate goal in reality is to lower the threshold for such technology to integrate into people's lives in the same way that people naturally interact with each other in a social sense. This approach may make it easier for some groups (such as older adult populations or school children) to make use of the technology, but it depends on the setting and intended use. We must think carefully about the *situated context of use* for different technologies. Indeed, what may make sense in one context may make zero sense another.

To put it another way, the interaction context is neither a blank canvas nor one monolithic construct. Rather, there are contextual boundaries that extend into the socio-technical systems we inhabit, and which are heavily influenced by cultural factors [61]. If one does something in Korea and the same thing in the United States, the response from others may be quite different, for a variety of reasons. Thus, while attempting to make things easier and more fluid for some groups, we may inadvertently be making things more difficult for other groups.

These challenges can also be expanded to apply to other multi-modal interaction contexts, such as conversational interfaces integrated into a wearable device or internet-of-things (IOT)-enabled HRI scenarios with interactive physical robots embedded into people's homes and work spaces [62]. Dealing with these broader societal impacts (and their potential challenges) will likely entail more research with diverse groups and settings in the future. Other approaches, such as participatory design, may also be critical in this for directly integrating user perspectives into our understanding of interaction components as another form of data [63]. Part of our motivation for creating the Social AI here as a multi-lingual system (capable of interacting in both English and Korean) is to explore some of these challenges in future work, but there are many potential avenues of research toward these issues.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/mti6020016/s1, Video S1: Social_AI_Brief_Example_Videos.mp4.

Author Contributions: Conceptualization, C.B. and B.W.; methodology, C.B. and B.W.; software, all; formal analysis, all; investigation, J.S., E.Y., J.J. and Y.C.; data curation, J.S., E.Y., J.J. and Y.C.; writing—original draft preparation, C.B. and B.W.; writing—review and editing, C.B. and B.W.; visualization, all; funding acquisition, C.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported through funding by a grant from the National Research Foundation of Korea (NRF) (Grant number: 2021R1G1A1003801).

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Hanyang University (protocol #HYU-2021-138). for studies involving humans.

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: The datasets generated during and/or analyzed during the current study are not publicly available due to the fact the data comprises video and audio recordings of identifiable human subjects during gameplay. However, extracted de-identified data may be made available from the corresponding author upon reasonable request.

Acknowledgments: We would also like to thank Cheda Stanojevic and Sawyer Collins (Indiana University) and Sungmin Yang (Hanyang University) for their assistance in this work.

Conflicts of Interest: The authors declare no conflict 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.

References

1. 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.
2. Gero, K.I.; Ashktorab, Z.; Dugan, C.; Pan, W.; Johnson, J.; Geyer, W.; Ruiz, M.; Miller, S.; Millen, D.R.; Campbell, M.; et al. Mental models of AI agents in a cooperative game setting. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI), Oahu, HI, USA, 25–30 April 2020; pp. 1–12.
3. Brennan, S.E. The grounding problem in conversations with and through computers. In *Social and Cognitive Psychological Approaches to Interpersonal Communication*; Fussell, S.R., Kreuz, R.J., Eds.; Lawrence Erlbaum: Hillsdale, NJ, USA, 1991; pp. 201–225.
4. Enfield, N. How we talk. In *The Inner Workings of Conversation*; BasicBooks: New York, NY, USA, 2017.
5. Koutsombogera, M.; Vogel, C. Speech pause patterns in collaborative dialogs. In *Innovations in Big Data Mining and Embedded Knowledge*; Esposito, A., Esposito, A.M., Jain, L., Eds.; Intelligent Systems Reference Library; Springer: Cham, Switzerland, 2019; pp. 99–115.
6. Knapp, M.; Hall, J. *Nonverbal Communication in Human Interaction*; Thomas Learning: Wadsworth, OH, USA; Boston, MA, USA, 2010.
7. Tseng, S.H.; Hsu, Y.H.; Chiang, Y.S.; Wu, T.Y.; Fu, L.C. Multi-human spatial social pattern understanding for a multi-modal robot through nonverbal social signals. In Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Edinburgh, Scotland, 25–29 August 2014; pp. 531–536.
8. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
9. Neff, G.; Nagy, P. Automation, algorithms, and politics | talking to Bots: Symbiotic agency and the case of Tay. *Int. J. Commun.* **2016**, *10*, 17.
10. Crosby, M. Building thinking machines by solving animal cognition tasks. *Minds Mach.* **2020**, *30*, 589–615. [CrossRef]
11. Honig, S.; Oron-Gilad, T. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Front. Psychol.* **2018**, *9*, 861. [CrossRef]
12. 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]
13. Doyle, P.R.; Clark, L.; Cowan, B.R. What do we see in them? identifying dimensions of partner models for speech interfaces using a psycholexical approach. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI), Yokohama, Japan, 8–13 May 2021; pp. 1–14.
14. Dennett, D. Intentional systems. *J. Philos.* **1971**, *68*, 87–106. [CrossRef]
15. Thomaz, A.; Hoffman, G.; Cakmak, M. Computational human-robot interaction. *Found. Trends Robot.* **2016**, *4*, 105–223.
16. Chesher, C.; Andreallo, F. Robotic faciality: The philosophy, science and art of robot faces. *Int. J. Soc. Robot.* **2021**, *13*, 83–96. [CrossRef]
17. Bennett, C.C.; Sabanovic, S.; Fraune, M.R.; Shaw, K. Context congruency and robotic facial expressions: Do effects on human perceptions vary across culture? In Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Edinburgh, Scotland, 25–29 August 2014; pp. 465–470.
18. Lisetti, C.L.; Brown, S.M.; Alvarez, K.; Marpaung, A.H. A social informatics approach to human-robot interaction with a service social robot. *IEEE Trans. Syst. Man Cybern. Part C* **2004**, *34*, 195–209. [CrossRef]
19. Muthugala, M.V.J.; Jayasekara, A.B.P. Enhancing user satisfaction by adapting Robot's perception of uncertain information based on environment and user feedback. *IEEE Access* **2017**, *5*, 26435–26447. [CrossRef]
20. 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]
21. 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 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), New York, NY, USA, 26–31 August 2016; pp. 507–512.
22. Fraune, M.R.; Oisted, B.C.; Sembrowski, C.E.; Gates, K.A.; Krupp, M.M.; Šabanović, S. Effects of robot-human versus robot-robot behavior and entitativity on anthropomorphism and willingness to interact. *Comput. Hum. Behav.* **2020**, *105*, 106220. [CrossRef]
23. Pearl, C. *Designing Voice User Interfaces: Principles of Conversational Experiences*; O'Reilly: Sebastopol, CA, USA, 2016.

24. Bernsen, N.O.; Dybkjaer, H.; Dybkjaer, L. *Designing Interactive Speech Systems. From First Ideas to User Testing*; Springer: New York, NY, USA, 1998.
25. International Telecommunication Union. *Parameters Describing the Interaction with Multimodal Dialogue Systems*; ITU-T Suppl. 25 to P-Series; International Telecommunication Union: Geneva, Switzerland, 2011.
26. Buisine, S.; Martin, J.C. The effects of speech-gesture cooperation in animated agents' behavior in multimedia presentations. *Interact. Comput.* **2007**, *19*, 484–493. [CrossRef]
27. Manson, J.H.; Bryant, G.A.; Gervais, M.M.; Kline, M.A. Convergence of speech rate in conversation predicts cooperation. *Evol. Hum. Behav.* **2013**, *34*, 419–426. [CrossRef]
28. Reitter, D.; Moore, J.D. Predicting success in dialogue. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 25–27 June 2007; pp. 808–815.
29. Lai, C.; Carletta, J.; Renals, S. Modelling participant affect in meetings with turn-taking features. In Proceedings of the Workshop of Affective Social Speech Signals, Grenoble, France, 22–23 August 2013.
30. Möller, S. *Quality of Telephone-Based Spoken Dialogue Systems*; Springer Science & Business Media: Bochum, Germany, 2004.
31. Munteanu, C.; Clark, L.; Cowan, B.; Schlögl, S.; Torres, M.I.; Edwards, J.; Murad, C.; Aylett, M.; Porcheron, M.; Candello, H.; et al. CUI: Conversational user interfaces: A workshop on new theoretical and methodological perspectives for researching speech-based conversational interactions. In Proceedings of the 25th International Conference on Intelligent User Interfaces Companion (IUI), Cagliari, Italy, 17–20 March 2020; pp. 15–16.
32. Anderson, A.H.; Bader, M.; Bard, E.G.; Boyle, E.; Doherty, G.; Garrod, S.; Isard, S.; Kowtko, J.; McAllister, J.; Miller, J.; et al. The HCRC map task corpus. *Lang Speech* **1991**, *34*, 351–366. [CrossRef]
33. Slater, M. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Phil. Trans. Biol. Sci.* **2009**, *364*, 3549–3557. [CrossRef]
34. Gonzalez-Franco, M.; Lanier, J. Model of illusions and virtual reality. *Front. Psychol.* **2017**, *8*, 1125. [CrossRef]
35. Edwards, C.; Edwards, A.; Stoll, B.; Lin, X.; Massey, N. Evaluations of an artificial intelligence instructor's voice: Social Identity Theory in human-robot interactions. *Comput. Hum. Behav.* **2019**, *90*, 357–362. [CrossRef]
36. Slater, M.; Antley, A.; Davison, A.; Swapp, D.; Guger, C.; Barker, C.; Pistrang, N.; Sanchez-Vives, M.V. A virtual reprise of the Stanley Milgram obedience experiments. *PLoS ONE* **2006**, *1*, e39. [CrossRef]
37. Rauchbauer, B.; Nazarian, B.; Bourhis, M.; Ochs, M.; Prévot, L.; Chaminade, T. Brain activity during reciprocal social interaction investigated using conversational robots as control condition. *Phil. Trans. Roy. Soc. Lond. B* **2019**, *374*, 20180033. [CrossRef]
38. Bennett, C.C. Evoking an intentional stance during human-agent social interaction: Appearances can be deceiving. In Proceedings of the 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), Vancouver, BC, Canada, 8–12 August 2021; pp. 362–368.
39. Hofer, M.; Hartmann, T.; Eden, A.; Ratan, R.; Hahn, L. The role of plausibility in the experience of spatial presence in virtual environments. *Front. Virtual Real.* **2020**, *1*, 2. [CrossRef]
40. Abdul, A.; Vermeulen, J.; Wang, D.; Lim, B.Y.; Kankanhalli, M. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI), Montreal, QC, Canada, 21–26 April 2018; pp. 1–18.
41. Kolokoltsov, V.N.; Malafeyev, O.A. *Understanding Game Theory: Introduction to the Analysis of Many Agent Systems with Competition and Cooperation*; World Scientific Publishing: Hackensack, NJ, USA, 2020.
42. Lim, S.; Reeves, B. Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player. *Int. J. Hum. Comput. Stud.* **2010**, *68*, 57–68. [CrossRef]
43. Bianchi, F.; Grimaldo, F.; Bravo, G.; Squazzoni, F. The peer review game: An agent-based model of scientists facing resource constraints and institutional pressures. *Scientometrics* **2018**, *116*, 1401–1420. [CrossRef] [PubMed]
44. Jesso, S.T.; Kennedy, W.G.; Wiese, E. Behavioral cues of humanness in complex environments: How people engage with human and artificially intelligent agents in a multiplayer videogame. *Front. Robot. AI* **2020**, *7*, 531805. [CrossRef] [PubMed]
45. Correia, F.; Alves-Oliveira, P.; Ribeiro, T.; Melo, F.; Paiva, A. A social robot as a card game player. In Proceedings of the 13th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, Salt Lake City, UT, USA, 5–9 October 2017; Volume 13, p. 1.
46. Völkel, S.T.; Schödel, R.; Buschek, D.; Stachl, C.; Winterhalter, V.; Bühner, M.; Hussmann, H. Developing a personality model for speech-based conversational agents using the psycholexical approach. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI), Oahu, HI, USA, 25–30 April 2020; pp. 1–14.
47. Nafcha, O.; Shamay-Tsoory, S.; Gabay, S. The sociality of social inhibition of return. *Cognition* **2020**, *195*, 104108. [CrossRef] [PubMed]
48. Klein, R.M.; MacInnes, W.J. Inhibition of return is a foraging facilitator in visual search. *Psychol. Sci.* **1999**, *10*, 346–352. [CrossRef]
49. Xia, K.; Huang, J.; Wang, H. LSTM-CNN architecture for human activity recognition. *IEEE Access* **2020**, *8*, 56855–56866. [CrossRef]
50. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
51. Siebert, J.; Joeckel, L.; Heidrich, J.; Nakamichi, K.; Ohashi, K.; Namba, I.; Yamamoto, R.; Aoyama, M. Towards guidelines for assessing qualities of machine learning systems. In Proceedings of the International Conference on the Quality of Information and Communications Technology (QUATIC), Online Conference, 8–11 September 2020; pp. 17–31.

52. Bennett, C.C.; Doub, T.W.; Selove, R. EHRs connect research and practice: Where predictive modeling, artificial intelligence, and clinical decision support intersect. *Health Policy Technol.* **2012**, *1*, 105–114. [CrossRef]
53. Ekman, P.; Friesen, W.V. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*; Malor Books: Los Altos, CA, USA, 2003.
54. Blom, P.M.; Bakkes, S.; Spronck, P. Modeling and adjusting in-game difficulty based on facial expression analysis. *Entertain. Comput.* **2019**, *31*, 100307. [CrossRef]
55. Mistry, K.; Jasekar, J.; Issac, B.; Zhang, L. Extended LBP based facial expression recognition system for adaptive AI agent behaviour. In Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–7.
56. Möller, S. Perceptual quality dimensions of spoken dialogue systems: A review and new experimental results. In Proceedings of the 4th European Congress on Acoustics (Forum Acusticum Budapest 2005), Budapest, Hungary, 29 August–2 September 2005; pp. 2681–2686.
57. Moore, R.K. From talking and listening robots to intelligent communicative machines. In *Robots That Talk and Listen*; Witz, J.M., Ed.; De Gruyter: Boston, MA USA, 2015.
58. Bowden, K.; Wu, J.; Oraby, S.; Misra, A.; Walker, M. SlugNERDS: A Named Entity Recognition Tool for Open Domain Dialogue Systems. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan, 7–12 May 2018.
59. Chakraborty, B.K.; Sarma, D.; Bhuyan, M.K.; MacDorman, K.F. Review of constraints on vision-based gesture recognition for human–computer interaction. *IET Comput. Vis.* **2018**, *12*, 3–15. [CrossRef]
60. Riek, L.D. Wizard of oz studies in HRI: A systematic review and new reporting guidelines. *J. Hum.-Robot. Interact.* **2012**, *1*, 119–136. [CrossRef]
61. Lee, H.R.; Šabanović, S. Culturally variable preferences for robot design and use in South Korea, Turkey, and the United States. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI), Bielefeld, Germany, 3–6 March 2014; pp. 17–24.
62. Anagnostis, A.; Benos, L.; Tsaopoulos, D.; Tagarakis, A.; Tsolakis, N.; Bochtis, D. Human activity recognition through recurrent neural networks for human–robot interaction in agriculture. *Appl. Sci.* **2021**, *11*, 2188. [CrossRef]
63. Lee, H.R.; Šabanović, S.; Chang, W.L.; Nagata, S.; Piatt., J.; Bennett, C.; Hakken, D. Steps toward participatory design of social robots: Mutual learning with older adults with depression. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI), Vienna, Austria, 6–9 March 2017; pp. 244–253.