*Multimodal Technologies and Interaction*

MDPI

*Article*

# Intelligent Blended Agents: Reality–Virtuality Interaction with Artificially Intelligent Embodied Virtual Humans

**Susanne Schmidt *, Oscar Ariza and Frank Steinicke**

Human-Computer Interaction, Department of Informatics, Universität Hamburg, 22527 Hamburg, Germany; ariza@informatik.uni-hamburg.de (O.A.); steinicke@informatik.uni-hamburg.de (F.S.)

* Correspondence: schmidt@informatik.uni-hamburg.de

Received: 19 October 2020; Accepted: 25 November 2020; Published: 27 November 2020


**Abstract:** Intelligent virtual agents (VAs) already support us in a variety of everyday tasks such as setting up appointments, monitoring our fitness, and organizing messages. Adding a humanoid body representation to these mostly voice-based VAs has enormous potential to enrich the human–agent communication process but, at the same time, raises expectations regarding the agent's social, spatial, and intelligent behavior. Embodied VAs may be perceived as less human-like if they, for example, do not return eye contact, or do not show a plausible collision behavior with the physical surroundings. In this article, we introduce a new model that extends human-to-human interaction to interaction with intelligent agents and covers different multi-modal and multi-sensory channels that are required to create believable embodied VAs. Theoretical considerations of the different aspects of human–agent interaction are complemented by implementation guidelines to support the practical development of such agents. In this context, we particularly emphasize one aspect that is distinctive of embodied agents, i.e., interaction with the physical world. Since previous studies indicated negative effects of implausible physical behavior of VAs, we were interested in the initial responses of users when interacting with a VA with virtual–physical capabilities for the first time. We conducted a pilot study to collect subjective feedback regarding two forms of virtual–physical interactions. Both were designed and implemented in preparation of the user study, and represent two different approaches to virtual–physical manipulations: (i) displacement of a robotic object, and (ii) writing on a physical sheet of paper with thermochromic ink. The qualitative results of the study indicate positive effects of agents with virtual–physical capabilities in terms of their perceived realism as well as evoked emotional responses of the users. We conclude with an outlook on possible future developments of different aspects of human–agent interaction in general and the physical simulation in particular.

**Keywords:** virtual agents; physical interaction; artificial intelligence; mixed reality environments

## 1. Introduction

Though personal digital assistants have become widespread in the context of smart homes as well as professional environments, most of the current implementations rely on audio output, displayed text or simple graphics only. Augmented reality (AR) technology can add a new dimension to existing agents by providing a humanoid 3D virtual body to complement the voice. Human-like AR representations can enrich the communicative channels that convey the agent's status and intentions to humans with gestures and other forms of social behaviors. Moreover, they can be registered spatially with their environment, which enables a more direct form of spatial interaction compared to voice-only interaction [1]. Several research projects addressed the question of whether and how agent embodiment affects the social interaction between virtual agents (VAs) and human communication partners.

In our previous work, we conducted two studies that we performed in a historical exhibition context to understand the importance of different humanoid representations of VAs [2,3]. We analyzed the effectiveness of virtual museum guides with varying embodiment (embodied vs. disembodied) with regard to several aspects, such as the user's feeling of co-presence, spatial presence, engagement, as well as the agent's credibility. The results indicate benefits of embodied audio-visual representations of VAs concerning all the before-mentioned metrics. These findings are in line with previous work summarized in Section 2.1.

With the majority of related literature attributing positive qualities to embodied agents as well as the rapid advancement of mixed reality (MR) technology, it seems reasonable to assume that users will have increasing access to services that include embodied VAs in the future. However, the inclusion of a body representation also poses new challenges to the interaction designers, in particular regarding the spatial and social integration of VAs in the physical environment. Since embodied VAs can be considered to be spatial, social entities, they should be subject to similar physical laws as well as social standards as human communication partners. We are therefore interested in the multi-sensory interaction between such agents and their real-world surroundings, including real interlocutors as well as physical objects.

Social integration of VAs requires the fusion of multiple information sources, including sensory cues, such as visual and auditory feedback, pre-existing knowledge, and information that is learned over the course of the interaction with users. In Section 3, we introduce a model that integrates concepts from the fields of computer vision, computer graphics, natural language processing, speech synthesis, and artificial intelligence (AI).

In this context, a challenging but important feature for VAs is the physically correct simulation of collisions between the agent and real-world-objects. Such a simulation requires complex actuators, in particular, if the technology behind the interaction should be hidden from the user to create an advanced illusion of plausible human-like agents. Consequently, only a few examples of virtual–physical interactions have been investigated so far; examples are discussed in Section 2.2. In the following, we term VAs with the capabilities to manipulate real-world objects as blended agents. In Section 4, we will evaluate two forms of such manipulations, i.e., hitting a physical ball and writing on a physical sheet of paper.

In this article, we contribute to the development of intelligent blended agents (IBAs) in the following ways:

1. Construction of a formal model to describe the multi-modal as well as multi-sensory interaction between IBAs and humans.
2. Formulation of guidelines to implement this model, exemplified on a suite of selected technologies.
3. Collection of subjective quantitative and qualitative user responses to investigate the ability of IBAs to manipulate physical objects.

The findings of the last-mentioned user study were originally reported in a conference paper [4]. In the present article, we extend our previous work by embedding the aspect of physical simulation into a comprehensive model of human–agent interaction.

## 2. Related Work

In the following, we will summarize previous research on the embodiment of VAs as well as agents with virtual–physical capabilities and intelligent behavior.

### *2.1. Embodied Virtual Agents*

A literature meta-review by Yee et al. [5] suggests that the inclusion of any visual representation of a VA's face leads to higher task performance measures. The presence of a virtual face seems to be more important than its visual quality. Therefore, even a representation with low realism can provide important social cues for human–agent interactions.

In two previous studies, we could show benefits of virtual embodied guides over those implemented just by voice in the context of museums [2,3]. Visitors of a simulated exhibition reported a higher sense of spatial and social presence in the company of embodied VAs. Furthermore, agent embodiment showed a significant effect on the perceived credibility of the VA, indicating that embodied guides appeared to be more competent and trustworthy.

A human-subject study by Kim et al. [6] showed similar positive effects of agent embodiment on the users' sense of trust, social richness, and social presence. In addition, participants of the study reported an increased confidence in the agent's ability to influence the real world and to react to real-world events, when the VA was embodied and showed natural social behaviors.

A literature review of early embodied VAs was presented by Dehn and van Mulken [7]. Their meta-analysis revealed some inconsistent findings regarding the effects of embodied VAs on user experience. Although some of the analyzed empirical studies report benefits of embodied VAs, others conclude that agent embodiment only showed little or even negative effects on the users' responses. Dehn and van Mulken hypothesize that these different outcomes may be attributed to varying degrees of the agent's appearance, both in terms of visual fidelity and natural voice.

In this context, the results of related studies indicate negative effects of a mismatch between visual and behavioral realism of embodied VAs on perceived co-presence [8] and perceived quality of communication [9]. Therefore, the visual representation of anthropomorphic VAs may raise expectations about their behavioral realism and should only be used if the system is able to meet these expectations [10].

### *2.2. Blended Agents*

Since embodied VAs represent spatial entities within their real-world environment, the question arises whether, and if so, to what extent they also must follow the same physical laws. Potential physicality conflicts include unnatural occlusions between the VA and physical objects, as well as implausible physical-virtual collisions.

Negative implications of such conflicts on human–agent interactions were demonstrated by Kim et al. [11,12]. In their studies, the participants observed a VA encountering a physical obstacle such as a door or a chair. In different conditions, the VA either avoided collisions with the obstacle, asked the participant to move it out of the way, or passed through the obstacle. Subjective responses indicate that physical-virtual conflicts reduced the sense of co-presence, while proactive behavior asking help from the users to avoid implausible conflicts increased the ratings of co-presence.

Instead of avoiding collisions between VAs and physical objects, another approach is to allow VAs to actually interact with these objects, for example, to move them to a different location. In this context, we introduced the concept of blended agents—VAs that are not only capable of influencing their virtual surroundings but also of performing virtual–physical interactions [4].

The concept of virtual–physical interactions was first investigated by Lee et al. [13]. They implemented a tabletop game that can be played by a real human and a VA. Via an actuator system underneath the surface of the table, the VA can move both virtual and physical tokens. In a within-subjects study, the authors were able to show benefits of the VA moving a physical token both regarding subjective and some behavioral measures. In this condition, participants reported a higher sense of co-presence and physicality, as well as higher expectations regarding the virtual human's abilities.

Lee et al. [14] demonstrated that even subtle tactile footstep vibrations induced via the floor can increase subjective estimates of co-presence in an AR environment.

Another example of VAs that can influence their physical environment was implemented by Lee et al. [15]. Their custom-made Wobbly Table crosses the boundary between the physical and virtual world. Although the physical half is standing in front of a projection screen, a virtual counterpart is visually extended into the virtual environment with the VA. If the VA is leaning on the table, a virtual–physical interaction occurs as both the virtual and the physical parts are slightly tilted.

In Section 4, we present and evaluate two prototypes of virtual–physical interactions using robotic actuators and thermochromic ink.

### *2.3. Intelligent Blended Agents*

In previous work, actions of blended agents were either completely predetermined or controlled by the experimenter to some extent, for example, by selecting the next dialog option or animation sequence. In both ways, some level of intelligence of blended agents can be simulated. With the rise of MR, machine learning (ML), and the Internet of things (IoT), blended agents can be further improved with AI. By this means, the agent can react to its current physical context, such as the user's behavior—without intervention of a human controller.

Consumer devices with embedded intelligent voice-controlled agents, such as Amazon Alexa or Apple's Siri, allow customers an intuitive form of interaction with their smart home environments and as a means to access information from the Internet [16]. By supplementing the intelligent agent with a body representation as well as virtual–physical capabilities, the resulting IBA could interact with its physical environment more naturally, using locomotion, gestures as well as other non-verbal cues. Kim et al. [6] present a human-subject study, in which users asked either an IBA or a voice-only agent to support them with certain tasks, such as turning off a light, checking the state of the lab, or closing a door. In comparison to the disembodied voice-only agent, the IBA was able to walk to the region of interest (i.e., the light switch, lab, or door) and perform gestures that reflect its current task. The authors hypothesize that an IBA can communicate its status and intentions more intuitively due to its body representation and intelligent social behavior. The results indicate that these factors indeed increase perceived social presence, as well as the user's confidence in the IBA's awareness of the real world and its ability to influence the physical environment.

Besides personal assistants that can visibly control connected smart home appliances, other forms of IBAs may be useful in different scenarios. For example, an IBA in a visitor information center could provide visitors with directions by drawing on a physical map. Also, an intelligent blended caregiver could support patients with performing physical training exercises.

## 3. Implementation of Intelligent Blended Agents

As we intend to provide a virtual 3D body representation for VAs that are embedded into a user's daily life (e.g., smart assistants or audio guides), we focus on interactions between an IBA and its real-world surroundings. This includes both real communication partners and physical objects. However, the model can be easily extended to additional virtual entities.

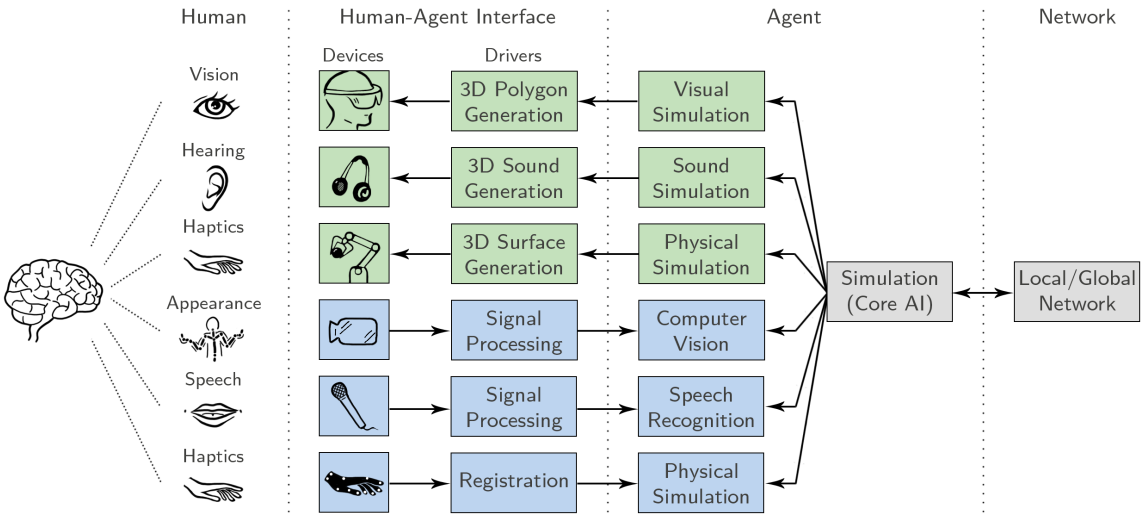



**Figure 1.** Model to illustrate multi-modal (blue boxes) multi-sensory (green boxes) interaction between a human and an IBA (adopted from Burdea and Coiffet [17]).

We base our considerations on the model of multi-modal as well as multi-sensory interaction in MR. This model describes interaction with a machine from a human's perspective using multiple input channels and output modalities. The machine processes the multi-modal (speech and haptic) output of the user and operates different simulation subsystems to produce visual, auditory, and haptic feedback. We adopt and modify this generalized model to describe human–agent interaction from the perspective of an IBA (see Figure 1). For this purpose, we extend the output modalities of the user by a third channel. This represents visual data an agent can collect and evaluate to gather more information about the current state of its physical environment, including the user. For the smart integration of the different user outputs and subsequent simulation of multi-sensory feedback, we introduce an *AI core* (cf. Section 3.6). The AI core can receive additional inputs from external knowledge bases in the local or global network.

In the following, each aspect of the model will be exemplified for our own implementation of an IBA, called Louise (see Figure 2). Implementation details are provided, allowing for replication by AR developers and researchers.

(**a**) (**b**)

**Figure 2.** Visual representation of (**a**) the agent's face, and (**b**) the agent's upper body.

### *3.1. Visual Simulation*

The embodiment of IBAs requires applying various techniques from the field of computer graphics to generate a 3D visual body representation. As discussed in Section 2.1, disparities between visual and behavioral realism should be decreased as they can result in lower levels of co-presence and social presence.

For a high degree of visual fidelity, we use a 3D scanned female head model that features a highly detailed mesh, 4K PBR textures, and multiple facial expressions (Animatable Digital Double of Louise by Eisko©, www.eisko.com). The head model, with slightly adapted textures and rendered using a screen-space subsurface scattering shader, is illustrated in Figure 2. To add vividness to Louise's face, its eyes are not moving randomly but are occasionally focusing on pre-defined points of interest, such as the user or specific tracked objects. Furthermore, micro movements, i.e., saccades, as well as blinking reflexes are performed. Lip movements that match Louise's voice are created via the Oculus Lip Sync plug-in in real time. Facial expressions representing different emotions are displayed using the model's blendshapes. As the blendshapes represent a variety of so-called action units, i.e., movements of facial muscle groups, complex emotions can be constructed using the Facial Action Coding System (FACS) [18]. For example, joy can be deconstructed into cheek raising and the lip corner pulling.

The agent's body was created and rigged using Adobe Fuse or Mixamo, respectively. Adobe Fuse allows adjusting several body parameters, including the height, proportions, and clothing of the agent. Based on the rigged 3D model, movement sequences can be either realized using retargeted

keyframe animations or motion-captured material, although the latter usually creates more natural results. For motion capturing, we use an 8-camera Qualisys Miqus M3 system and a full-body mocap suit. All animation sequences are performed by a human actor, recorded and exported to the FBX file format using the Qualisys Track Manager (QTM), and imported to Unity. Since there is an automated mapping between QTM segments and Unity bones, animations can be applied directly to a humanoid rig that was set up using the Unity animation system Mecanim. Differences in the proportions between the captured actor and the rig can be solved using an additional inverse kinematics (IK) pass in Unity. For example, the function *Foot IK* reduces feet sliding by slightly adjusting the skeleton pose in a way that the agent's feet hit the same spots on the ground as in the original animation of the actor.

### *3.2. Sound Simulation*

The audio output of Louise is synthesized using the IBM Watson text-to-speech engine (https://www.ibm.com/cloud/watson-text-to-speech). This service offers voice models with varying language, gender, and tone. For Louise, we use the female American English voice LisaV3, which relies on Deep Neural Networks that are pre-trained on human speech data [19]. If the regular pronunciation rules of the specified language do not apply to a set of domain-specific words, a custom dictionary can be created for these exceptional cases. In Unity, such a dictionary consists of key-value pairs, i.e., a custom word and the corresponding pronunciation. The latter can be specified either as a sounds-like translation (e.g., *IEEE* sounds like *I triple E*) or a phonetic translation (e.g., aɪ ˈtrɪpəl i). To make use of the Watson text-to-speech API, the Unity application must be authenticated using a unique API key. The service credentials, including this key as well as an endpoint URL for the selected data center, can be accessed from the IBM cloud web interface after creating a new service instance (https://cloud.ibm.com/resources). This procedure is identical for all Watson services that are used for Louise and can be simplified by using the Watson SDK for Unity (https://github.com/watson-developer-cloud/unity-sdk). After authentication, a JSON file containing the text to synthesize, as well as optional values for the requested audio format, voice, and custom voice model, is sent to the Watson text-to-speech service. By default, the audio is synthesized with a sampling rate of 20,050 Hz and is streamed back to the Unity client as an array of bytes. Optionally, additional word timings can be requested as an alternative to synchronize lip movements with the spoken text (in contrast, our current solution solely relies on the audio data as described in the last section). Sending text to the IBM cloud data center and receiving synthesized audio adds a delay that can depend on several factors such as the client and server locations as well as available bandwidth. To reduce the delay, we store generated audio files in a local cache for future use. The cache is limited to a user-defined size and is managed according to the least recently used (LRU) policies. In Unity, received audio files can be bound dynamically to an audio source that is registered in 3D space to match the current pose of Louise. To improve the realism of sound propagation, we use the Unity MS HRTF spatializer plug-in, which incorporates the binaural head-related transfer function.

### *3.3. Physical Simulation*

One main purpose of simulating the somatosensory system is to create physically plausible reactions to collisions between the embodied agent and entities in the real-world environment. The approximation of virtual–physical collisions requires an accurate registration between real scene objects and the agent. This could be realized based on spatial information that is collected through (visual) tracking approaches, as discussed in Section 3.4. Results of object tracking can be directly fed back to the agent's animation system, for example, to implement grasping movements that target the tracked physical object. In Unity, we use IK to naturally transform the joints of the IBA's skeleton based on the target poses of its hands, fingers, or feet. To not only align the agent's body with objects, but also to manipulate their physical properties, more complex actuators are necessary (as introduced in Section 2.2).

For this purpose, we implemented two different setups that exemplify virtual–physical interactions in the form of (i) movements of a physical ball, and (ii) persistent writing on a physical sheet of paper. To address the former, we used an off-the-shelf robotic golf ball that moves along a scripted path to simulate interaction with a virtual golf club (our custom JavaScript library for the robotic ball can be found at https://github.com/augmentedrealist/spheromini.js). In other scenarios, robotic devices could not only be used to represent manipulable physical entities but also to actuate otherwise passive real-world objects. For the second form of virtual–physical interactions, we designed a novel device that uses temperature variation to activate thermochromic ink on a sheet of paper. In the current implementation of the thermal table as a proof of concept, all written text must be prepared before the MR experience. Only when placed on the powered table, the prepared text turns invisible until the polarity of the thermoelectric element is changed, and the sheet of paper is cooled down. Synchronized with the animations of a blended agent, the illusion of a virtual human writing on a physical piece of paper can be created. Further technical details for both the thermal table and the robotic ball can be found in [4].

Besides the simulation of collisions between the IBA and physical objects, other somatosensory functions could be integrated using additional sensors and actuators. For example, Kim et al. [20] used a wind sensor to detect airflow from a real fan and to adjust the agent's behavior accordingly.

### *3.4. Computer Vision*

Non-verbal cues, such as facial expressions, gestures, or eye contact, are crucial in human-to-human communication. Therefore, they should also be incorporated into human–agent interactions to make them as natural as possible. By leveraging computer vision algorithms, at least some of those cues can be recognized and interpreted based on plain RGB camera input. For Louise, we focus on the identification of users based on their facial features as well as the extraction of their eye positions to make eye contact. Although there is a Watson visual recognition API that could be used for both tasks, it is not optimized for the distinction of faces and therefore, may result in a lower accuracy than a dedicated library. This is because, in general, all faces are similar to respect to certain features, such as the presence of eyes, a nose, and a mouth, as well as rough proportions. Custom approaches, such as FaceNet [21], use a deep convolutional network to compute a feature vector per image that can be directly used to compute face similarities between two images. We set up a dedicated Python server that runs a Tensorflow implementation of FaceNet (https://github.com/davidsandberg/facenet). In preparation of the face recognition, an initial training procedure was performed. We first collected five images for each user that should be known to Louise (the best results are achieved when they are captured within the same environment as for the final application). The images are aligned using a multi-task cascaded convolutional neural network (MTCNN). In this step, faces within the images are detected, cropped, and scaled to an image size of $160 \times 160$ pixels with the result that eyes and the mouth appear at similar positions in all images. For each aligned image, a 128-dimensional feature vector, a so-called embedding, is computed using a pre-trained model (Model 20180402-114759, which has been trained on the VGGFace2 dataset). Based on these embeddings, we built a classifier using a scikit-learn implementation of support vector machines [22]. The alignment step and the construction of a classifier took 27.888 and 14.458 s, respectively, when performed on an Intel Core i7-7660U processor without GPU acceleration. After the initial training procedure, the Python server can receive and classify frames of a video stream that are sent by the Unity client via HTTP POST requests. In Unity, a WebCam Texture is used to display the live camera feed. For requesting a face classification, the current texture content is encoded into PNG format and sent to the Python server as a byte array. The server's response includes whether a face was detected and, if so, the associated user with a confidence measure. If the confidence measure is below a threshold, it is assumed that none of the users in the database were identified. For our prototype, we pre-trained 20 users; however, the classifier could also be updated with new training data while the Unity application is running.

An even more accurate approximation of human vision could incorporate additional depth sensors to infer spatial relationships within the physical surroundings. We use two approaches: (i) an optical tracking system (5 OptiTrack Prime 13W cameras) to gather 6 DOF data of tracked objects or users, and (ii) a system of depth cameras (3 Microsoft Kinect v2) if a full 3D point cloud of the scene is necessary. The captured depth data can be used to adjust the agent's behavior, for example, by avoiding a user's motion path, making eye contact, or grasping an object. Furthermore, if the depth of different body parts of the user is known, the user's skeletal pose can be derived. By this means, the agent can react to gestures of the user.

### *3.5. Speech Recognition*

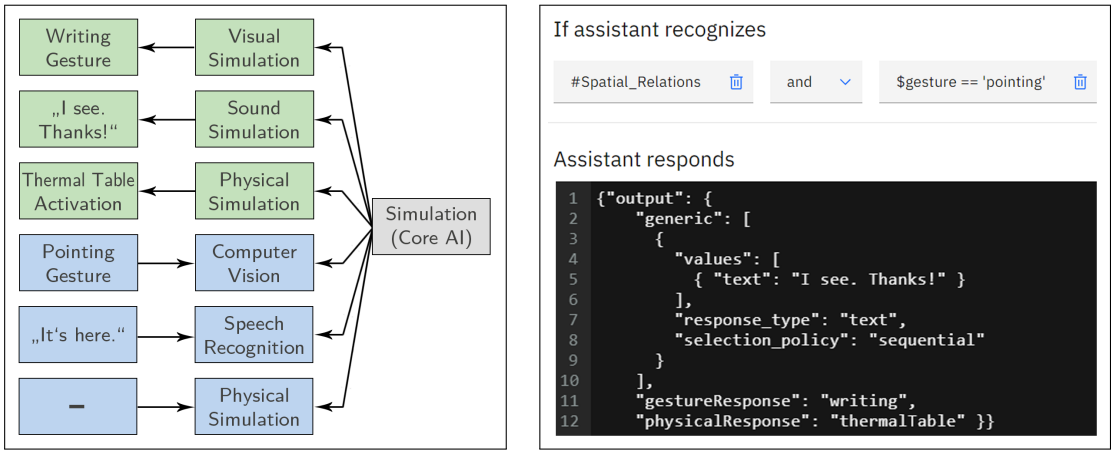
To add another layer of awareness to Louise, we are recording the user's voice and transfer it to the IBM Watson speech-to-text service (https://www.ibm.com/cloud/watson-speech-to-text) for further processing. The Watson service is based on deep neural networks and allows the transcription of speech input in various languages [23]. It incorporates knowledge of the structure and grammar of a selected language, combined with an acoustic model that describes the composition of voice signals. If the agent is expected to handle specific vocabulary, for example, in the context of a computer science application, a domain-specific language model that augments the existing base model can be trained. Training requires a set of individual words or a domain-specific text from which the system can extract unknown terms. For the former, a pronunciation must be added manually along with each word. In contrast, the system can identify words from their context when provided within full training sentences, and therefore, no pronunciation is needed in this case.

To make sure that the agent does not react to its own voice, speech recognition is paused as long as the agent is talking. After the audio input was processed, a JSON message is sent to the Unity application, including the final transcript with a confidence measure. The transcribed speech can then be further processed, as presented in the following section.

### *3.6. Core AI*

The planning and execution of Louise's actions, including speech output, movements, and facial expressions, require a fusion of different sources of information. These can include multiple sensory channels (visual, auditory, and somatosensory), a pre-populated database with basic knowledge of the agent, an artificial memory that is dynamically updated, as well as external sources (e.g., a calendar or weather app). All collected information is integrated using the IBM Watson Assistant (https://www.ibm.com/cloud/watson-assistant). Key component of the assistant is a dialog skill that must be trained before it is used for real-time human–agent interaction. Training data is collected for three subsystems: intents, entities, and the dialog tree. An intent is an objective a user might pursue, interpreted from the text input the assistant gets. To recognize user intents without depending on the specific phrasing, sample utterances are provided to the system that reflect possible text inputs representing the same intent. Entities can provide additional context to an intent. For example, a general room reservation intent could be specified by adding entity values for the requesting user, the room to be reserved, as well as a date. As for intents, examples can be provided by the developer to allow for automated recognition of an entity within the text input. The dialog tree integrates all provided information to decide on the agent's output. It constitutes the initial knowledge base of the agent since all possible reactions to a specific input within a current context are pre-defined. Although the Watson Assistant was developed to carry on conversations with users depending on recognized utterances, it can be extended to non-textual input by receiving additional information from Unity. The exchange of messages between Unity and Watson is based on JSON files. A JSON request sent by the Unity client can contain a user's utterance that was transcribed by the speech recognition system as well as custom variables. The latter can represent other sensory inputs (e.g., the visually identified user), data recalled from the agent's memory (e.g., the last time a user was seen), and information retrieved from external sources (e.g., calendar events on a specific date). Custom variables can be stored in

Watson context variables for later use. For an incoming message, all child nodes of the current dialog node are sequentially processed until the condition of one of them evaluates to true. Conditions can consider the message itself, regarding the occurrence of a specific intent, entity, or entity value, as well as the stored context. When a dialog node is triggered, the Watson assistant returns the associated JSON response to the Unity application. The response can include any recognized intents or entities, each with a confidence score, a text output as well as self-defined variables. Examples of the latter include an associated emotion, a matching animation, or a description of additional media files. All JSON responses within the Watson Assistant dialog tree are fully customizable. Therefore, it is the developers' responsibility to maintain the consistency between different modalities, for example, by specifying matching verbal and behavioral output for a specific dialog node. After received by the Unity client, the JSON response is interpreted, and further actions in the subsystems of visual, sound, and physical simulation are initiated. Overall, the core AI's task is to match multi-modal inputs provided by the Unity application to multi-sensory responses of the IBA, while incorporating the agent's current state of knowledge. Figure 3 shows an exemplified interaction with audio and gesture input as well as audio, gesture, and physical output. The illustrated scenario is further detailed in the following section.




(**a**) (**b**)

**Figure 3.** (**a**) Exemplified multi-modal multi-sensory interaction between a human and an IBA, and (**b**) corresponding representation within the IBM Watson Assistant dialog tree. The symbols # and $ are denoting intents and context variables, respectively.

## 4. User Study: Virtual–Physical Manipulations

For implementing the model of human–agent interaction that is presented in the previous section, we can mostly draw on existing services such as IBM Watson and FaceNet. The aspect of physical simulation, however, is still an active subject of investigation since only a few practical solutions for virtual–physical manipulations in MR environments are available. We proposed two systems to implement different forms of virtual–physical interactions—one similar to previously tested techniques and one that features a novel approach. In this section, we present a comparative study that we conducted to collect subjective responses of users. For this purpose, we designed an experimental environment that naturally embeds both implementations, as described in detail in the following subsections.

### *4.1. Hypothesis*

Throughout the following sections, we focus on two different forms of virtual–physical interactions, i.e., manipulations of (i) an object's location, and (ii) an object's material properties. Based on these two approaches, we formulate the following hypothesis:

**Hypothesis 1 (H1).** *Virtual–physical interactions improve the user experience in terms of social and spatial presence, ecological validity, perceived anthropomorphism of the blended agent, and engagement.*

**Hypothesis 2 (H2).** *Virtual–physical interactions related to the surface material of an object have a stronger positive impact on the aforementioned metrics than those related to the object's position.*

**Hypothesis 3 (H3).** *Chemical changes of an object's surface material can be hidden from the user, while mechanical manipulations of the object's location are more explicable.*

**Hypothesis 4 (H4).** *Manipulations of the object's location are observed by the users directly, while manipulations of the object's surface material are not observed before the end of the MR experience.*

(H3) and (H4) are assumed to have a direct effect on (H2). This is because the illusion of an interaction between the blended agent and the physical object might be supported if users are not capable of finding an obvious explanation for an effect. Furthermore, persistent changes of an object's surface material are observable even after the MR experience and, therefore, could change the perceived realism of the VA in retrospect.

### *4.2. Materials*

The MR environment for the user study was inspired by a minigolf course, with a VA acting as the opposing player. For the two forms of virtual–physical manipulations, we used a robotic golf ball and a thermal table, as summarized in Section 3.3 and described in detail in [4]. The course with a length of 2.9 m and a width of 1.05 m was set up within a four-sided CAVE, which featured a size of $4.2 \times 3.13 \times 2.36$ m. Two heavy ropes marked the edges of the course, and pieces of artificial turf served as obstacles. The hole was marked with a slightly raised ring. Three Optoma EH320USTi short throw projectors were used to display virtual content at the CAVE walls. Two additional Optoma GT1080(e) projectors augmented the floor as well as the thermal table. Each projector provided a resolution of $1920 \times 1080$ at a refresh rate of 120 Hz. To experience the view-dependent stereoscopic content, users had to wear shutter glasses with passive markers that were tracked by a five-camera OptiTrack system. Furthermore, the voice of the VAs was presented to participants via wireless noise-cancelling headphones. Another purpose of the headphones was to block ambient noise that was created by the thermal table's internal fan. In contrast, sounds caused by the friction between the golf ball and the floor were still audible. All the actions of the participants were monitored by the experimenter using a camera at the ceiling of the CAVE. By this means, the experimenter was able to trigger particular reactions of the blended agent from a neighboring room, without being visible to the participants. The CAVE setup was chosen since current AR headsets only feature a small field of view, and virtual objects are cut off at the edges of the display area.

For the VA, we used the rigged model as described in Section 3.1, with adapted clothing, hairstyle, and makeup. As retargeted keyframe animations were described as stiff and artificial in a pre-study, we decided to replace them with motion-captured animation sequences that were performed by a female actor. A path of the robotic golf ball was programmed accordingly to match the animations of the blended agent. The agent's voice was provided by a native speaker and spatialized in Unity, as described in Section 3.2. Besides the voice, no additional sound effects such as footsteps were simulated in the current experiment. To better control for extraneous influences and increase the repeatability of the user study, we used a wizard of oz approach to time reactions of the agent and to select correct dialog options. By this means, we could ensure the same sequence of events for study participants who were experiencing different conditions. In a real application, this could be replaced by the modules that are described in Sections 3.4–3.6.

### *4.3. Methods*

For the user study, we followed a between-subjects design with two independent variables and two levels each. The resulting four conditions, all in relation to the VA's interactions, are:

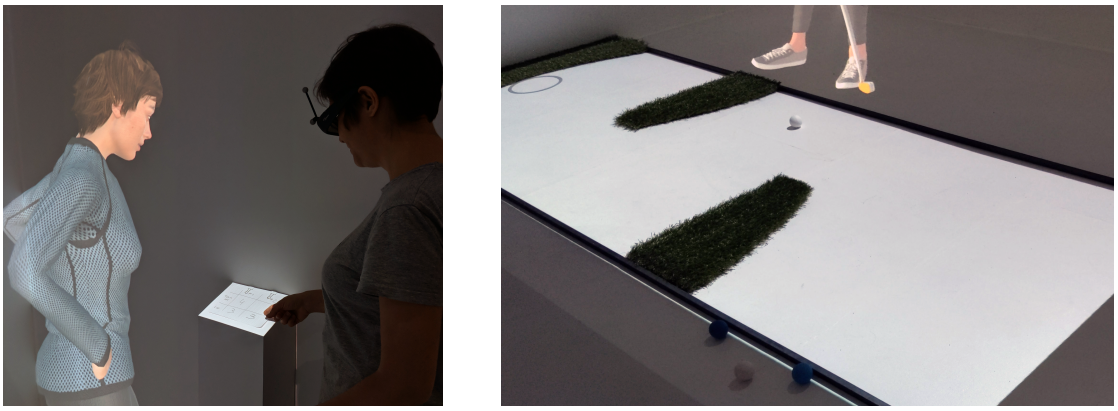
- ($B_vH_v$) Virtual golf ball and virtual handwriting.
- ($B_vH_r$) Virtual golf ball and real handwriting.
- ($B_rH_v$) Real golf ball and virtual handwriting.
- ($B_rH_r$) Real golf ball and real handwriting.

The random assignment of conditions was counterbalanced among participants. Before a new participant arrived, the golf course was prepared according to the selected condition. In the conditions ($B_vH_r$) and ($B_rH_r$), the thermal table was turned on, and a sheet of paper prepared with invisible ink was placed on its top. In preparation for the conditions ($B_rH_v$) and ($B_rH_r$), a NodeJS server was started, which connected Unity to the robotic ball. Since the robotic ball is not able to provide a global orientation value, its initial rotation had to be determined by hand. A manual correction was performed until the ball moved perfectly along a test track. Afterward, the ball was positioned at its starting slot along with three other physical golf balls.

Before they entered the previously described minigolf course, participants had to fill in a consent form as well as a pre-questionnaire to provide demographic information. Afterward, each participant was guided to the CAVE, and the procedure, as well as general minigolf rules, were explained. In this introductory phase, participants were able to examine the golf course as well as the scorecard with their naked eye. Also, the preparation of the scorecard with a table was executed in sight of the participants to make sure that they realize it was empty when they entered the room. After all questions were resolved, the participant had to wear shutter glasses, and the experimenter started with the first round.

In total, four rounds of play were performed: (1) The experimenter playing with a physical ball, (2) the participant playing with a physical ball, (3) the (blended) agent playing with a virtual/robotic ball, and (4) the participant playing with a physical ball. After rounds (1) and (2), the experimenter and participant filled in one blank of the scorecard each. Afterward, participants were introduced to the VA by the experimenter. They were given noise-cancelling headphones, and the experimenter left the room. The VA then started a conversation with the participant and putted either a virtual or a robotic ball into the hole, according to the selected condition (see Figure 4b). After finishing round (3), the VA walked to the scorecard and asked the participant where to fill in her score, as illustrated in Figure 4a. The animation was only resumed by the experimenter if the participant was close to the table. This artificial pause should ensure that all participants witnessed the handwriting of the VA and, therefore, do not make false assumptions about how and when the VA's score was added to the scorecard. If the participant was in sight of the thermal table, the VA virtually wrote a pre-defined score of 4 in the dedicated blank space. The scenario is illustrated in Figure 3a with a possible Watson implementation in Figure 3b (though, a wizard of oz approach was used in the experiment as motivated in Section 4.2). The VA then challenged the participant to bet her score in round (4). During this second round of the participant, all hits were counted by the experimenter using the live camera view. In the conditions ($B_vH_r$) and ($B_rH_r$), when the participant was close to the hole, the temperature of the thermal table was switched from high to low, and the thermochromic ink below the projected score became visible. At the same moment, the projected score was faded out with the result that the participant could only see the physically written score when he returned to the table.

After round (4) was finished, and the participant filled in the last blank of the table, the VA started a final evaluation of the match. Based on the number of strokes that were digitally logged by the experimenter, the VA announced the winner of the game. Finally, the VA said goodbye and suggested to the participant to take the scorecard as a souvenir. The user left the CAVE and was asked to fill in a series of questionnaires. Overall, the study took around 20 to 25 min to complete.



(**a**) (**b**)

**Figure 4.** Photos showing the experimental setup, with (**a**) the thermal table, and (**b**) the minigolf course with a robotic ball.

### *4.4. Participants*

We invited 40 participants to our study, 27 male and 13 female (aged from 18 to 41, $M = 25.35$). 36 of them were students or staff members of the local department of informatics, while 4 stated to pursue a non-technical profession. According to the pre-questionnaire, 7 participants took part in a study involving MR for the first time. None of the 40 participants reported any visual impairments that could affect the results of our experiment.

### *4.5. Results*

During the user study, we collected both quantitative and qualitative subjective data that can give some indication of how different virtual–physical interactions affect a MR experience. The results are presented in the following section.

#### 4.5.1. Quantitative Analysis

Each experiment session was concluded with five questionnaires that addressed different aspects of the experience:

- Social presence (=Social Presence Questionnaire by Bailenson et al. [24])
- Spatial presence (=subscale of the Temple Presence Inventory [25])
- Ecological validity (=subscale of the ITC Sense of Presence Inventory [26])
- Perceived anthropomorphism (=subscale of the Godspeed questionnaire [27])
- Engagement (=top-loading items of the engagement subscale of the ITC Sense of Presence Inventory [26])

Results were measured on 7/5-point Likert scales, as noted in the head of Table 1. For each participant, average scores were formed according to the computation models that are suggested in the original papers. We analyzed the data using multiple two-way ANOVAs, but could not find any significant main or interaction effects. The mean values, as well as standard deviations for all conditions and each dependent variable, are also summarized in Table 1. Additional analyses of demographic factors, such as gender, age, and experience with MR studies, did not indicate any moderate or strong correlations with the dependent variables.

**Table 1.** Mean scores and standard deviations for each of the 4 conditions and 5 dependent variables.

| | | **Social Presence (1–7)** | | **Spatial Presence (1–7)** | | **Ecological Validity (1–5)** | | **Anthropo-Morphism (1–5)** | | **Engagement (1–5)** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Golf ball | Writing | M | SD | M | SD | M | SD | M | SD | M | SD |
| virtual | virtual | 4.74 | 1.139 | 4.600 | 0.824 | 3.840 | 0.506 | 3.620 | 0.745 | 4.567 | 0.522 |
| virtual | real | 4.20 | 1.178 | 4.643 | 0.678 | 3.520 | 0.738 | 3.320 | 0.694 | 4.233 | 0.545 |
| real | virtual | 4.80 | 0.660 | 4.671 | 1.048 | 3.720 | 0.634 | 3.320 | 0.655 | 4.500 | 0.503 |
| real | real | 4.16 | 0.970 | 4.586 | 0.995 | 3.700 | 0.620 | 3.260 | 0.795 | 4.467 | 0.477 |

#### 4.5.2. Qualitative Analysis

In addition to the mentioned Likert scales, we were also asking participants some open-ended questions about experienced or expected effects of blended agents, depending on the tested condition:

($B_r$) "How did the agent's interaction with a real golf ball affect your experience?"
($B_v$) "Imagine the agent interacting with a real golf ball instead of a virtual one. How would this interaction have affected your experience?"
($H_r$) "How did the agent's persistent handwriting affect your experience?"
($H_v$) "How would your experience have been changed, when the handwritten score of the agent would be still visible on the paper?"

We directed similar questions at participants assigned to the virtual and the real conditions as we were interested in the users' perception of individual potentialities of both virtual and blended agents. By this means, we also aimed to investigate whether expectations for virtual–physical interactions and the reality diverge to some extent. To extract comparable data, we assigned utterances to seven different categories using an open coding strategy. The first three categories denote opinions regarding the perceived realism of blended agents in comparison to VAs without virtual–physical capabilities. Another four categories cover different dimensions of emotional responses. If a single response of a participant included multiple utterances within the same category (e.g., "fascinating and memorable"), they were still counted only once. As each scenario ($B_r$, $B_v$, $H_r$, and $H_v$) was experienced by 20 participants, 20 is the maximum value in each category. The resulting frequency distribution is illustrated in Table 2.

**Table 2.** Categorization of the users' utterances in open-ended questions related to the user experience.

| | **Sub-Category** | **Example** | **Count** | | **Count** | |
|---|---|---|---|---|---|---|
| | | | $B_r$ | $B_v$ | $H_r$ | $H_v$ |
| Realism | More realistic | 'It made me feel like I am playing against a real human.' | 6 | 10 | 7 | 5 |
| | Less realistic | 'It looked a little unrealistic how the golf club was hitting the ball.' | 2 | 1 | 0 | 0 |
| | No difference | 'I don't think that it would have changed much.' | 2 | 4 | 3 | 4 |
| Emotions | Disconcertment | 'Would've probably felt more weird.' | 3 | 5 | 2 | 3 |
| | Confusion | 'Initially there was a bit of confusion whether it is the real ball.' | 4 | 2 | 3 | 2 |
| | Surprise | 'It was a fun surprise to see the actual ball be moved.' | 7 | 0 | 6 | 3 |
| | Enjoyment | 'It made the experience more fascinating and memorable.' | 11 | 3 | 11 | 7 |

In addition to the categorized utterances, four participants acknowledged the increased fairness when both the user and the blended agent must play with a physical golf ball. Regarding the scorecard, four of the participants with a ($H_v$) condition reported their initial surprise when the projected score

disappeared. One participant even admitted feeling disappointed as he could not take the completed scorecard as a souvenir. In general, three participants mentioned that a scorecard with physically written text feels more like a trophy.

We were also interested in whether reactions to the persistent handwriting were different for users assigned to a ($B_r$) or ($B_v$) condition. Experiencing a blended agent that is interacting with a real golf ball might have raised the expectations regarding the agent's capabilities. However, no such interaction effect could be found as both user groups showed similar, mainly positive, responses to the persistent handwriting.

Finally, we asked participants of the ($B_r$) or ($H_r$) conditions about the used mechanism to test our hypothesis (H3). For the physical golf ball, 8 of the participants stated that they figured out the mechanism or at least got an idea of how it worked. Surprisingly, only one of the ideas was correct while most participants suspected a magnetic track behind the ball movement. In contrast, none of the ($H_r$) participants perceived the mechanism behind persistent handwriting as obvious. Due to a lack of explanations, two participants were convinced that another person entered the room to replace the virtual score, while another two felt uncertain about the fact whether the scorecard was empty at the beginning of the study.

#### 4.5.3. Observational Data

In addition to feedback obtained through the questionnaires, we also made some observations regarding the participants' behavior, both during the study and directly afterward.

When the agent was approaching the scorecard to fill in the blank, six of the participants (five of $H_r$ and one of $H_v$) used their own pen to write down the agent's score in her stead.

After the MR experience but before completing the post-questionnaires, all participants of the conditions ($B_rH_v$), ($B_vH_r$), and ($B_rH_r$) were asked whether they noticed the physicality of the golf ball and handwriting, respectively. 16 participants who experienced a $B_r$ condition realized that the ball was real during the experiment, while 4 reported having doubts whether the ball was real or not. In contrast, only 7 participants of the $H_r$ conditions noticed that the handwritten score was still persistent after they left the MR environment. These results support our initial hypothesis (H4) regarding the observability of both forms of virtual–physical interactions. It was surprising, though, as we expected that users will realize the persistent handwriting at the latest when they leave the MR environment.

After completion of the entire experiment, 20% of participants with one of the $H_v$ conditions took their scorecard home. In the $H_r$ conditions featuring persistent handwriting, 40% of participants kept their scorecard.

### *4.6. Discussion*

Against our hypotheses (H1) and (H2), no significant effects of physical simulation on several social and spatial metrics could be found in the collected data. These statistical results can be interpreted in different ways. Two possible implications might be that (i) there actually are no differences between VAs without virtual–physical capabilities and blended agents, and (ii) some users react positively to blended agents while others show negative reactions, compensating one another. In contradiction to both approaches to an explanation are the qualitative comments that were collected at the end of the study. Most participants reported a positive influence of both virtual–physical interactions in terms of perceived realism and/or user experience. The question arises why the quantitative ratings do not reflect these subjective impressions. In the following, we discuss several potential influencing factors and rate their respective impact.

#### 4.6.1. Limited Expectations of VAs

Starting point of the following discussion is the basic question "Do users expect VAs to have physical capabilities?". The fact that the majority of participants did not notice the persistent handwriting at all and showed emotional responses of surprise and confusion when they were

made aware of it is indicative of rather low expectations of the VA. Therefore, users assigned to a virtual condition most likely compared the displayed VA to their mental model of agents, without a negative impact of missing physical capabilities. This model may be shaped either by previous first-hand experiences with VAs or depictions of such agents in the media, including movies and books. This impression is supported by the qualitative feedback as users of the virtual condition felt positive about the realistic body movements, the natural voice, and individual reactions of the VA. Therefore, the reported effects found in previous within-subjects studies might be only due to the direct comparison between agents.

#### 4.6.2. Low Granularity of Used Scales

Three of the questionnaires used 5-point Likert scales as we complied with the standards. As current VAs are still far from being indistinguishable from real humans, only a few users will rate items that are related to the human likeness with a maximum value of 5. Therefore, only one of the remaining options refers to a positive response. A higher scale granularity that allows participants to rate their experience more precisely might reveal significant differences between the conditions. That these differences are expected to be rather small was already indicated by the results of a similar study with a within-subjects design [13]. For the ratings of engagement, an additional ceiling effect can be observed as mean scores are already close to the maximum for conditions without blended agents.

#### 4.6.3. Limited Importance of the Physical Reactions

Although both the golf ball and the scorecard were designed to be an integral part of the interaction between the participants and the MR environment, they might have been less meaningful than other interactions with physical objects. For example, if a blended agent moves a real chair towards the user to take a seat, this physical manipulation has an impact on the subsequent actions while the physical golf ball could only be observed without any direct contact. In another example, a blended agent could mark a location on a physical map to direct the user to a place. In contrast, the scorecard was only given as a souvenir without any future purpose. More meaningful interactions might have increased the perceived value of the physical (persistent) manipulations.

#### 4.6.4. Distrust of the Experimental Environment

An observation that might have influenced the results without being the sole reason was that some participants conjectured that somebody entered the room and replaced the virtual score by a physical one when the participant was distracted by the golf match. Even the fact that the experimenter was neither in the CAVE nor the directly neighboring room could convince them of the contrary. Two other participants mentioned that they were sure that the sheet of paper was empty at first but were skeptical about this in retrospect as they would not know how this could have been done.

### *4.7. Limitations*

Both the interactions with the robotic ball and the thermal table are proof of concepts with some limitations.

As the robotic ball has no global tracking capabilities, its position and orientation must be determined manually. A computer vision algorithm could be used to compute the current position of the ball and to match the subsequent actions of the blended agent. Such a tracking algorithm could also compensate for the limited precision of the robotic ball. In the current implementation, the ball ends up in slightly varying places. Although the blended agent always putted the ball into the hole, the golf club and the ball movements were not always perfectly in sync; an observation that was also shared by some of the participants. Another limitation is related to the ball physics. As any motorized objects need an acceleration phase to be set in motion, the initial impulse imparted to the golf ball by hitting it with a golf club cannot be simulated completely.

The current implementation of the thermal table also requires some preparations to create a convincing illusion of a blended agent. First, the ink can only be made visible at once. Therefore, the handwritten text must be prepared before the MR experience. For the same reason, the writing path must be simulated virtually before it is replaced by the physical writing. To solve both problems, coated paper could be used. In this case, the thermal mechanism must be changed from a heating plate to a heated metal tip as used for soldering irons. Furthermore, as the used thermochromic ink is visible at room temperature, it always must be placed on top of the thermal table at the beginning of a MR experience. By using a dye with different characteristics users could bring a sheet of paper to the MR environment, which might further increase the believability of the blended agent.

## 5. Conclusions

In this article, we introduced a model for the interaction between humans and intelligent blended agents (IBAs). Such agents are characterized by their capabilities to (i) receive multi-modal feedback from one or more users, (ii) process this input data using various AI approaches and with respect to prior knowledge, and (iii) generate plausible multi-sensory outputs, i.e., behavior that can be perceived by users based on their visual, auditory, and haptic channels. In particular, the generation of haptic output is challenging, as it requires IBAs to manipulate their real-world surroundings.

In the second part of the article, we provide insights into this emerging field of research by presenting a user study, which compares two forms of such virtual–physical interactions (moving a physical golf ball and writing on a physical sheet of paper). Although a statistical analysis of subjective data obtained through several questionnaires did not yield any significant differences between agents with and without physical capabilities, user responses still indicate benefits of the former. Users described their interaction with blended agents as an "amazing, very surprising and immersive experience", a "fascinating magic trick", or the sensation of "being inside the Holodeck". The agent's physical manipulations "made the agent appear more present", and created a "more enjoyable" and "more memorable" MR experience. Despite little divergences between the natural and simulated behavior of a physical golf ball, most participants appreciated the virtual–physical interaction. Some participants also mentioned effects on their behavior inside the MR environment, as they "felt the urge to respond" to the blended agent or avoided any collisions.

In a future study, we plan to address the aspects that were identified in the discussion, including the consideration of more meaningful virtual–physical interactions as well as the usage of questionnaire scales with a higher granularity together with objective feedback. Since the observed emotional responses of participants could be primarily attributed to the novelty effect, we are particularly interested in changes in user behavior when interacting with blended agents. By this means, more valuable indications for the long-term use of VAs could be inferred.

Regarding the introduced model of human-agent interaction, each component can be extended by using advanced processing or simulation techniques. Our Unity framework for embodied IBAs is structured in a modular way, with the result that individual functions can be easily modified or even replaced. Although we based multiple modules such as the speech recognition and generation as well as the AI dialog system on IBM Watson services, other vendors such as Google Cloud offer APIs with similar functionality. Since dependencies between the Unity modules are well defined, they can be individually exchanged according to the developer's needs and preferences.

The core AI, which is implemented as a capsuled module itself, can also be extended, for example, to incorporate additional internal or external data sources. In the current prototype, reactions of our IBA Louise are generated with respect to different input channels and a knowledge base that represents her cognitive intelligence. Since a major purpose of IBAs is to interact socially with human users, a useful extension of this model is to also include aspects of artificial emotional intelligence. Louise can already show complex facial expressions; however, she is not capable of recognizing emotional states of the user so far. This could be realized by adding functionality to the image and speech processing

modules (e.g., by using the Watson tone analyzer). Endowing Louise with an artificial empathy towards human communication partners is one of the future research directions we plan to pursue.

**Author Contributions:** Conceptualization, S.S.; Methodology, S.S.; Software, S.S. and O.A.; Validation, S.S.; Formal Analysis, S.S. and O.A.; Investigation, S.S.; Resources, F.S.; Data Curation, S.S.; Writing—Original Draft Preparation, S.S.; Writing—Review & Editing, F.S. and O.A.; Visualization, S.S.; Supervision, F.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Alibali, M. Gesture in Spatial Cognition: Expressing, Communicating, and Thinking About Spatial Information. *Spat. Cogn. Comput.* **2005**, *5*, 307–331. [CrossRef]
2. Schmidt, S.; Bruder, G.; Steinicke, F. Effects of Embodiment on Generic and Content-Specific Intelligent Virtual Agents as Exhibition Guides. In Proceedings of the International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (ICAT-EGVE), Limassol, Cyprus, 7–9 November 2018; pp. 13–20.
3. Schmidt, S.; Bruder, G.; Steinicke, F. Effects of Virtual Agent and Object Representation on Experiencing Exhibited Artifacts. *Elsevier Comput. Graph.* **2019**, *83*, 1–10. [CrossRef]
4. Schmidt, S.; Ariza, O.; Steinicke, F. Blended Agents: Manipulation of Physical Objects within Mixed Reality Environments and Beyond. In Proceedings of the ACM Symposium on Spatial User Interaction (SUI), New Orleans, LA, USA, 19–20 October 2019; pp. 1–10.
5. Yee, N.; Bailenson, J.; Rickertsen, K. A Meta-Analysis of the Impact of the Inclusion and Realism of Human-Like Faces on User Experiences in Interfaces. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), San Jose, CA, USA, 28 April–3 May 2007; pp. 1–10.
6. Kim, K.; Boelling, L.; Haesler, S.; Bailenson, J.; Bruder, G.; Welch, G. Does a Digital Assistant Need a Body? The Influence of Visual Embodiment and Social Behavior on the Perception of Intelligent Virtual Agents in AR. In Proceedings of the IEEE Symposium on Mixed and Augmented Reality (ISMAR), Munich, Germany, 16–20 October 2018; pp. 105–114.
7. Dehn, D.; van Mulken, S. The Impact of Animated Interface Agents: A Review of Empirical Research. *Int. J. Hum. Comput. Stud.* **2000**, *52*, 1–22. [CrossRef]
8. Bailenson, J.; Swinth, K.; Hoyt, C.; Persky, S.; Dimov, A.; Blascovich, J. The Independent and Interactive Effects of Embodied-Agent Appearance and Behavior on Self-Report, Cognitive, and Behavioral Markers of Copresence in Immersive Virtual Environments. *Presence Teleoper. Virtual Environ.* **2005**, *14*, 379–393. [CrossRef]
9. Garau, M.; Slater, M.; Vinayagamoorthy, V.; Brogni, A.; Steed, A.; Sasse, M. The Impact of Avatar Realism and Eye Gaze Control on Perceived Quality of Communication in a Shared Immersive Virtual Environment. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), Fort Lauderdale, FL, USA, 5–10 April 2003; pp. 529–536.
10. Nowak, K.; Biocca, F. The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments. *Presence Teleoper. Virtual Environ.* **2003**, *12*, 481–494. [CrossRef]
11. Kim, K.; Maloney, D.; Bruder, G.; Bailenson, J.; Welch, G. The Effects of Virtual Human's Spatial and Behavioral Coherence with Physical Objects on Social Presence in AR. *Comput. Animat. Virtual Worlds* **2017**, *28*, e1771. [CrossRef]
12. Kim, K.; Bruder, G.; Welch, G. Exploring the Effects of Observed Physicality Conflicts on Real–Virtual Human Interaction in Augmented Reality. In Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST), Gothenburg, Sweden, 8–10 November 2017; pp. 1–7.
13. Lee, M.; Norouzi, N.; Bruder, G.; Wisniewski, P.; Welch, G. The Physical-Virtual Table: Exploring the Effects of a Virtual Human's Physical Influence on Social Interaction. In Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST), Tokyo, Japan, 28 November–1 December 2018; p. 25.

14. Lee, M.; Bruder, G.; Welch, G. Exploring the Effect of Vibrotactile Feedback Through the Floor on Social Presence in an Immersive Virtual Environment. In Proceedings of the IEEE Conference on Virtual Reality (VR), Los Angeles, CA, USA, 18–22 March 2017; pp. 105–111.
15. Lee, M.; Kim, K.; Daher, S.; Raij, A.; Schubert, R.; Bailenson, J.; Welch, G. The Wobbly Table: Increased Social Presence via Subtle Incidental Movement of a Real-Virtual Table. In Proceedings of the IEEE Conference on Virtual Reality (VR), Greenville, SC, USA, 19–23 March 2016; pp. 11–17.
16. Kim, K.; Billinghurst, M.; Bruder, G.; Duh, H.B.L.; Welch, G. Revisiting Trends in Augmented Reality Research: A Review of the 2nd Decade of ISMAR (2008–2017). *IEEE Trans. Vis. Comput. Graph.* **2018**, *24*, 2947–2962. [CrossRef] [PubMed]
17. Burdea, G.; Coiffet, P. *Virtual Reality Technology*; John Wiley & Sons: Hoboken, NJ, USA, 2003.
18. Ekman, P.; Rosenberg, E. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*; Oxford University Press: New York, NY, USA, 1997.
19. Kons, Z.; Shechtman, S.; Sorin, A.; Rabinovitz, C.; Hoory, R. High Quality, Lightweight and Adaptable TTS Using LPCNet. *arXiv* **2019**, arXiv:1905.00590.
20. Kim, K.; Bruder, G.; Welch, G. Blowing in the Wind: Increasing Copresence with a Virtual Human via Airflow Influence in Augmented Reality. In Proceedings of the International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (ICAT-EGVE), Limassol, Cyprus, 7–9 November 2018; pp. 183–190.
21. Schroff, F.; Kalenichenko, D.; Philbin, J. FaceNet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 815–823.
22. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
23. Kurata, G.; Ramabhadran, B.; Saon, G.; Sethy, A. Language Modeling with Highway LSTM. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Okinawa, Japan, 16–20 December 2017; pp. 244–251.
24. Bailenson, J.; Blascovich, J.; Beall, A.; Loomis, J. Interpersonal Distance in Immersive Virtual Environments. *Personal. Soc. Psychol. Bull.* **2003**, *29*, 819–833. [CrossRef] [PubMed]
25. Lombard, M.; Ditton, T.; Weinstein, L. Measuring Presence: The Temple Presence Inventory. In Proceedings of the International Workshop on Presence (PRESENCE), Los Angeles, CA, USA, 11–13 November 2009; pp. 1–15.
26. Lessiter, J.; Freeman, J.; Keogh, E.; Davidoff, J. A Cross-Media Presence Questionnaire: The ITC-Sense of Presence Inventory. *Presence Teleoper. Virtual Environ.* **2001**, *10*, 282–297. [CrossRef]
27. 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]