Computers in Human Behavior 141 (2023) 107606

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Vero: An accessible method for studying human–AI teamwork☆

Aaron Schecter [a,*], Jess Hohenstein [b], Lindsay Larson [c], Alexa Harris [d], Tsung-Yu Hou [b], Wen-Ying Lee [b], Nina Lauharatanahirun [e], Leslie DeChurch [d], Noshir Contractor [d], Malte Jung [b]

[a] University of Georgia, Athens, GA, 30602, United States
[b] Cornell University, Ithaca, NY, 14853, United States
[c] University of North Carolina, Chapel Hill, NC, 27599, United States
[d] Northwestern University, Evanston, IL, 60208, United States
[e] Pennsylvania State University, State College, PA, 16801, United States

ARTICLE INFO

Keywords:
Human computer interaction
Human-Agent Teaming
Experimental methods
Wizard of Oz
Remote collaboration

ABSTRACT

Despite the recognized need to prepare for a future of human–AI collaboration, the technical skills necessary to develop and deploy AI systems are considerable, making such research difficult to perform without specialized knowledge. To make human–AI collaboration research more accessible, we developed a novel experimental method that combines a standard video conferencing platform, a set of animations, and Wizard of Oz methods to simulate a group interaction with an AI teammate. Through a case study, we demonstrate the flexibility and ease of deployment of this approach. We also provide evidence that the method creates a highly believable experience of interacting with an AI agent. By detailing this method, we hope that researchers regardless of background can replicate it to more easily answer questions that will inform the design and development of future human–AI collaboration technologies.

1. Introduction

Smart technology is everywhere; we drive autonomous vehicles, talk to virtual assistants, play games in virtual reality, and receive feedback from our appliances. With improvements in automation technology, autonomous agents – often powered by artificial intelligence (AI) – are taking on more complex roles in their interactions with people and, consequentially, are increasingly viewed as more than mere tools (Jung & Hinds, 2018; Schaefer, Straub, Chen, Putney, & Evans III, 2017; Sebo, Stoll, Scassellati, & Jung, 2020). This greater integration is particularly salient in the domain of group work, with many companies implementing human-agent teams (HATs) in the workplace. While there is speculation that HATs might function differently than traditional teams, there has not been a significant amount of empirical research examining these differences. More extensive research is therefore needed to understand how these autonomous teammates change the nature of work. For the purpose of this paper, we will refer to "agents" as autonomous entities created to perform some aspect of teamwork independently of humans. Though these agents can take many forms (e.g., robots, chatbots), we assume the agent has some embodiment (a visual form) and can have some verbal interaction with a human.

The introduction of autonomous teammates invites new theorizing on how the processes and outcomes of human-agent teams extend or amend current theories on human teams. Specifically, it is unclear how these agents can regulate or change team interaction. Further, research is needed to identify the ways autonomous agents might mitigate or heighten the need for particular team processes. Answering such questions about the impact of autonomous agents on how humans interact, think, and feel is a critical but understudied area.

Among the many identified open questions and topics of interest (Rzepka & Berger, 2018; Seeber et al., 2018), there has been progress in the development of scales (Moussawi & Koufaris, 2019) and frameworks (Dellermann et al., 2021) to measure human perceptions of autonomous agents and describe the ways the use them to complete tasks. There have also been some empirical studies of human-agent teams (Jung & Hinds, 2018; Jung, Martelaro, & Hinds, 2015; Sebo et al., 2020; Traeger, Sebo, Jung, Scassellati, & Christakis, 2020). However, current research is somewhat limited in that many researchers do not have access to autonomous agents, or do not have the technical capacity to develop them. In general, researchers that do not have access to autonomous agents are lacking the platforms and methods that enable them to not only understand how such systems affect work in teams but also to inform future development of such agents. Even research approaches that rely on simulated AI agents often requite a significant amount of technical expertise. Limiting the ability to do

☆ This research is supported by the US Army Research Laboratory, award number W911NF-19-S-0001.

* Corresponding author.
E-mail address: aschecter@uga.edu (A. Schecter).

https://doi.org/10.1016/j.chb.2022.107606
Received 31 May 2022; Received in revised form 15 November 2022; Accepted 11 December 2022
Available online 13 December 2022

research on autonomous agents to those with technical knowledge or access to complex technical resources limits scholarly discourse with implications for theorizing and future technology development.

The growing prevalence of human-agent collaboration presents a need for experimental methods that allow researchers from diverse academic communities to more easily study complex social interactions with autonomous agents. Experiments are a critical tool for studying HATs due to the numerous variables at play in such complex settings and the difficulty in isolating them in practice. We argue that the field as a whole would benefit from a widely accessible and robust methodology for conducting such experiments.

In this paper, we present a unique experimental method that has allowed us to examine user interactions with an autonomous agent teammate using a convenient, accessible, and easy-to-use video conferencing application (Zoom). We refer to this method as the *Vero method*. The name is derived from the initial moniker we gave the AI agent created for our experiments. The Vero method makes human-agent collaboration research accessible to a broad community by eliminating the need to develop an autonomous agent and instead using only a widely available video conferencing platform. We make the materials[1] available for researchers to run their own experiments, and thus provide a standardized approach for conducting HAT studies. This standardized approach could also speed up experimental research and foster more rapid knowledge creation in this domain. Further, we make HAT studies more feasible for many researchers by removing the need to re-establish experimental protocols. Because we rely upon existing Wizard-of-Oz methods, the Vero method has the potential to increase the scalability and generalizability of human-agent experimental studies. Our use of a video-conferencing platform allows researchers to conduct multiple simultaneous study sessions and our approach allows the use of natural spoken language rather than pre-recorded speech or text-to-speech generation methods. This feature provides a great degree of flexibility for researchers, given that many tasks are compatible with the Vero method. We demonstrate through a case study how the method can be effectively deployed in an experimental team task. From our experiences with the case study we provide recommendations for future researchers deploying their own studies on human-agent teaming.

2. Related work

2.1. Research on human-agent teams

Technology is integrated into teams in a variety of ways. One way to consider the integration of intelligent technology onto teams is through the levels of automation (Parasuraman, 2000) or levels of autonomy (O'Neill, McNeese, Barron, & Schelble, 2022) of the technology. At the simplest level, where there is little to no autonomy of the technology, technology may provide a means for members of a team to communicate (e.g., using Zoom to video-conference). At more complex levels, technology plays a focal role in how teams complete their tasks, or serves as a tool the team controls (e.g., a drone, or robotic assisted surgery). At the most complex level of autonomy, the smart technology is itself an agentic member of the team. As the capabilities of autonomous systems continue to increase, AI agents are transitioning towards this highest level, acting less as tools and more like teammates capable of making independent and team-oriented decisions (Schaefer et al., 2017). Technologies acting in a teammate role will likely require shifts in team states and processes, as well as the ways in which leaders lead their human-agent teams (Larson & DeChurch, 2020). Such agents could be especially useful teammates in high-risk situations where human lives are at stake, such as military, space flight, and emergency response teams.

The highest degree of machine integration into a team is when the technology itself is an autonomous member of the group (Jung & Hinds, 2018). Here, the technology is able to both interact with group members and make decisions independent of the human team members. Thus, technology is no longer merely a tool, but instead a peer that actively works towards the team goals. For instance, a human may team with a robot to collaboratively search an area during a search and rescue mission (Jung et al., 2013). In this scenario, the human relies on the robot to use its sensors to detect threats and then communicate its assessment. The human then decides what to do based on the robot's advice. The unique aspect of this case is that the machine agent does not require human input to carry out "unpredictable" elements of its task. Thus, the machine not only improves the efficiency of human team members, but also makes decisions independently of its human counterparts in potentially unpredictable ways. Importantly, the machine agent may also have the capacity to learn from and adapt to the human members of the team.

Although many people will not likely interact with such complex agents any time soon, less intrusive forms of AI agents are already becoming increasingly prevalent. For example, "smart replies", which are already widely-used in various communication applications (e.g., Gmail), consist of an AI agent that suggests responses for users based on the conversation context (Kannan et al., 2016). Recent work has found that even this relatively minimal agent interference alters users' language and interpersonal perceptions (Hohenstein & Jung, 2018, 2020), suggesting the importance of research that investigates the social implications of new forms of AI. As AI agents become increasingly involved in team interactions, it is critical to investigate how such technologies affect team processes and how we can develop agents that capitalize on group dynamics and technological capabilities while avoiding potential perils.

In the human–AI collaboration literature, some of the most pressing open questions involve user perceptions of AI humanness, capabilities, and transparency (Rzepka & Berger, 2018; Seeber et al., 2018). Researchers within the field of information systems have started to investigate such questions, with recent work examining topics ranging from symbiotic co-evolution of human–AI teams (Döppner, Derckx, & Schoder, 2019), trust of intelligent systems (De Visser et al., 2020; McNeese, Demir, Chiou, Cooke, & Yanikian, 2019; You & Robert, 2018) and interaction design (Bittner & Shoury, 2019; Dellermann et al., 2021; Derrick & Elson, 2019; Dolata, Kilic, & Schwabe, 2019) to developing scales for measuring perceived AI intelligence and anthropomorphism (Moussawi & Koufaris, 2019).

There is increasing interest in understanding how machines such as robots impact the social processes of groups and teams. A recent literature review has highlighted the impact a robot's behavior can have on the dynamics of groups and teams (Sebo et al., 2020) and researchers have begun to theorize a robot's role within a team (e.g., Abrams & der Pütten, 2020). Other studies have examined how integrating an autonomous agent with multiple humans can produce a variety of adaptations in how individuals do work and relate to others. For instance, recent work has demonstrated that humans can form an emotional attachment to robotic teammates, and this emotional attachment can subsequently improve team performance and viability (You & Robert, 2017). Some social machines are designed explicitly to modify or improve human behaviors in a group setting. Yet other research shows how robots can be designed to mediate interpersonal interaction (Traeger et al., 2020). For example, the addition of a robot to a human dyad can improve conflict resolution dynamics (Jung et al., 2015). Even machines that are not intentionally designed to shape interpersonal interactions can do so by reshaping social norms (Lee, Kiesler, Forlizzi, & Rybski, 2012).

[1] Materials include animation files, an introductory video, debriefing script, and confederate training materials. These are detailed in a README file. Files and an FAQ board can be found at https://github.com/Robots-in-Groups-Lab/Vero-Method.

As a whole, this work highlights the pressing need to build detailed understanding about the impact of complex autonomous machines on groups and teams. With this technology increasingly shaping how we work, it is particularly important to allow researchers from a broad set of disciplines to participate in developing this understanding and in influencing the design of such systems.

2.2. Contemporary methods for studying human-agent teams

There are a variety of methods for studying human-agent teams, including case studies, vignettes, and experiments. In case-based research, a physical robot is typically embedded in a real-world situation, and then researchers record humans' reactions to the agent. Studies include employee responses to an automated snack delivery robot (Lee et al., 2012), behavioral changes in response to an automated delivery robot in a hospital (Mutlu & Forlizzi, 2008), adaptations to a manufacturing robot (Sauppé & Mutlu, 2015), reactions to service robots in shopping malls or airports (Kanda, Shiomi, Miyashita, Ishiguro, & Hagita, 2009; Triebel et al., 2016), and adjustments to hospital team routines when using a surgical robot (Beane, 2019; Cheatle, Pelikan, Jung, & Jackson, 2019; Pelikan, Cheatle, Jung, & Jackson, 2018; Sergeeva, Faraj, & Huysman, 2020). These studies have provided key evidence for how humans react to an autonomous agent and adapt interactions with other humans because of the presence of the autonomous agent. However, such studies can be prohibitive because they require a physical robot, which can be expensive or outside the technical capabilities of a research team.

To overcome this limitation, some researchers have used a vignette methodology where experimenters provide a scenario to participants in which they imagine themselves as an actor within the scenario and respond to survey questions accordingly. Vignette studies allow researchers to study human–AI collaboration in a more accessible manner when the technologies of study interest are inaccessible to researchers because of factors such as technical design, costs, or participant usability, for example. Another benefit of vignette studies is relatively simple implementation of study manipulations by simply changing the wording of vignettes for different manipulation conditions of interest, as has been done in some AI transparency work (De Fine Licht, Naurin, Esaiasson, & Gilljam, 2014). For example, in a study on AI decision-making transparency, Yu and Li (2022) presented a vignette where the participant was working in a human–AI collaboration with an AI system and other humans in an automobile company and the AI was the primary decision-maker in the tasks. Experimenters changed pieces of the vignette scenario in order to manipulate AI decision-making transparency. Another benefit of vignettes is that participants can be more easily exposed to multiple conditions. For example, Lima, Grgić-Hlača, and Cha (2021) developed vignette scenarios adapted from real-life events of AI-assisted bail decision-making to study how people attribute moral responsibility to AI decision-making. In another study looking at human-agent collaboration between clinicians and machine learning recommender systems in medical treatment selection, participants were shown hypothetical patient scenario vignettes (Jacobs et al., 2021). Notably, there are important limitations in using a vignette methodology, especially in the study of experiences as novel as human-agent collaboration. In particular, Yu an Li acknowledge that, because of their vignette methodology, they were not able to "fully elicit the true psychological reaction of the participants with this method, which limited the external validity" (p. 13). Further, Jacobs and colleagues also acknowledge that their vignette methodology represents a limitation to their research and thus future work "should also examine these models in real-world clinical workflows" (2021, p. 8).

Beyond case studies or vignettes, researchers can run experimental sessions with small teams using a more traditional battery of tasks from the psychology and organizational behavior literatures. There are several exemplar studies using experiments to study human–AI collaboration. In one study, small teams worked to cooperatively solve a puzzle with a small mobile robot (Jung et al., 2015), and the robot took actions to diffuse or enhance conflict. Other similar studies tested the effect of vulnerable language use by a robotic teammate (Traeger et al., 2020) and the development of emotional attachment between a human and their robotic teammate (You & Robert, 2017). Other researchers have attempted to circumvent these issues through simulated environments (Lee et al., 2021; Pynadath, Wang, Rovira, & Barnes, 2018; Wang, Pynadath, Rovira, Barnes, & Hill, 2018; Wong et al., 2021). In general, experiments are a popular means of studying HATs, though researchers must deal with challenges of cost, scale, and standardization.

In this study, we build on these methods by designing a standardized experimental approach that makes studying human-agent teaming viable for a wide variety of researchers. Specifically, we seek to maintain the realism of case studies, the accessibility of vignettes, and the validity of experiments. To ensure researchers can collect sufficient data – and expand the participant pool beyond the easily accessible – we develop a remote video-conferencing paradigm in line with similar work on HAT and HRI (Feil-Seifer, Haring, Rossi, Wagner, & Williams, 2020; Lematta et al., 2022). While our study is not the first to introduce this type of technique (e.g., Lematta et al., 2022), we focus on designing and validating a paradigm that is easy to use, customizable, and accessible to researchers seeking to study HATs even when they do not have a technology background.

2.3. Overcoming the challenges to studying human-AI collaboration

2.3.1. Wizard-of-Oz methods

Wizard of Oz ("WoZ") methods originated in the early 1980s (Green & Wei-Haas, 1985; Kelley, 1983) as a "testing or iterative design methodology wherein an experimenter (the 'Wizard'), in a laboratory setting, simulates the behavior of a theoretical intelligent computer application" (Kelley, 2018). The term "wizard of Oz" is used in reference to a novel by Frank Baum titled the "The Wonderful Wizard of Oz" in which an old man pretends to be a powerful wizard by operating props from behind a screen.

Although the approach originally focused on the design and testing of novel natural language based human–computer interfaces at a time when natural language processing capabilities were highly limited, it has since developed into a general approach to simulate intelligence for automated systems through involvement of a human-wizard. WoZ methods are typically used when a specific technology is either not yet available (e.g. robots with certain social capabilities), not attainable by a research team (e.g. a proprietary new language processing algorithm), or when it is important to learn about the impact of a technology before taking the effort to implement it (e.g. a design team wants to explore the viability of an idea before investing development resources). For these reasons, WoZ has become a mainstay in research and design of human–robot interaction (Riek, 2012), and of interactions with intelligent agents (Bittner & Shoury, 2019; Derrick & Elson, 2019; McNeese et al., 2019). With this wide application, the wizard's purpose has developed beyond natural language processing to include decision making (e.g., Jung et al., 2020) or the enactment of specific motion characteristics.

More recently, WoZ approaches have developed from an approach that was focused on testing a specific intelligent system or a precisely specified behavior of an intelligent system, into a general design technique for exploratory prototyping that allows researchers to discover and develop novel system behaviors on the fly (Zamfirescu-Pereira et al., 2021). For example, several studies by Sirkin and Ju have used WoZ methods to explore a wide range of behaviors and interaction patterns for simple, low-degree of freedom robots (Sirkin, Mok, Yang, & Ju, 2015). Perhaps most importantly, the scalability of WoZ methods has recently been improved with the advent of remote experimental platforms (Lematta et al., 2022). These advances, in part due to the COVID-19 pandemic, have made it more feasible to recruit sufficient participants and maintain experimental validity.

2.3.2. *Video conferencing as a research platform*

Video conferencing as a research tool has become increasingly popular due to its relatively low cost (Deakin & Wakefield, 2014; Sedgwick & Spiers, 2009), ability to access larger numbers of more diverse participants (Deakin & Wakefield, 2014; Sedgwick & Spiers, 2009), elimination of the need for participants to travel, efficiency, and ability to reduce various unpredictable circumstances (Sedgwick & Spiers, 2009). In previous research examining Zoom as a qualitative interviewing platform, participants reported having a positive experience and enjoyed the convenience of and time saved by not having to physically go to a lab, ease of using a platform that they were already familiar with, and accessibility via a range of platforms (i.e., phone, tablet, computer). In addition to the benefits for participants, the researchers noted that using Zoom made it economically feasible to recruit a large number of diverse, geographically-distributed participants. Given the noted prevalence of under-powered studies in psychological research and with most study designs and analyses requiring hundreds of participants at a minimum (Brysbaert, 2019), tools that allow researchers to easily run internet-based studies and experiments are increasingly important.

While the present study leverages Zoom, a variety of other platforms are viable choices and share many of the same advantages. Platforms such as Microsoft Teams and GoToMeeting also do not require participants to make an account or download a program to use it. Zoom includes password protection for confidentiality and the ability to record sessions directly to the host’s computer, increasing privacy. Other programs are also highly secure, in particular Teams which has stringent cyber security and privacy policies. A further benefit of video conferencing technologies is that many, including Zoom, save recorded sessions as both an audio file and a combined audio–video file, making any post-processing (e.g., transcription, video image analysis) more manageable. Further, Zoom produces (imperfect) transcripts of the audio files. In general, while we use one platform for our study, researchers should be able to use a variety of videoconferencing platforms to carry out their studies given the array of accessibility and secure features available.

Using Zoom as our research platform and employing WoZ methods, we created an experimental method that allows us to study interactions between an AI agent and one or more human team members without the need to develop a functional AI agent. This method allows researchers to investigate relevant questions about the design and development of future AI agent technologies *without* expending the resources needed to create such agents. Our novel methodology couples a WoZ method with videoconferencing to create a scalable approach by simulating multiple simultaneous interactions with an AI agent. This method enables researchers to recruit from any population with internet access, run multiple simultaneous trials, allow participants to take part without coming to a physical lab, allow participants more meaningful/valid human–AI interactions, and rapidly alter AI agent characteristics of interest.

3. The Vero method

We introduce Vero as a novel, accessible method to simulate an AI teammate. This approach was developed in response to the difficulties of running in-person studies during the COVID-19 pandemic. Inspired by previous research that explored videoconferencing as a research platform (e.g., Brodsky, Lee, & Leonard, 2021; Feil-Seifer et al., 2020; Sedgwick & Spiers, 2009; ?) and by other research using WoZ methods (e.g., Bittner & Shoury, 2019; Derrick & Elson, 2019; McNeese et al., 2019), we used an iterative design approach (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007) to develop a method that would allow us to simulate an AI-teammate with a setup that is accessible to researchers without any programming skills. We initially gave the teammate the moniker *Vero*; since then we have evolved to calling the approach the *Vero Method*. Below we describe the key components of the Vero Method and how it can be applied. To make its application easier, we also provide all necessary materials as Supplementary Files.

3.1. *Apparatus and materials*

The Vero Method relies on three components: A video conferencing platform that allows the use of virtual animated backgrounds (e.g. Zoom), a set of animations for each of the non-verbal behaviors the agent can perform, and an introduction video that establishes the belief in participants that artificial agents that act as teammates and can converse in ways indistinguishable from people are a technical possibility.

3.1.1. *Video conferencing system*

The Vero Method relies on a video conferencing tool that allows the use of virtual animated backgrounds. We used Zoom for the purposes of our study. Other helpful features include background noise suppression to filter out sounds that could destroy the illusion of an intelligent agent and reveal the WoZ nature of the setup (e.g. typing sounds, car noise, or barking dogs).

The complete video conferencing setup for studies includes several accounts: one account for each Vero, one account for each human participant, and, if breakout groups are used to run studies, an additional “recorder” account is needed to capture a video-recording of the interactions (at the time this research was performed, Zoom did not allow recording of breakout groups).

3.1.2. *Vero animations*

Fig. 1 depicts the five animations used in the Vero Method. Vero is an avatar representation of an animated intelligent agent that was created in Blender (Blender Foundation, 0000). Blender is a free, open-source 3D rendering software that can be used to create a variety of animations, such as the ones that we created to represent Vero. We created five different animations to highlight the agent’s states and support conversational functions: (1) a slow bouncing motion mimicking a breathing pattern to indicate an idling state, (2) a quick contracting motion used for backchanneling to indicate listening and attention (3) an outwards radiating motion to indicate a speaking state (4) a jumping motion to indicate the intention to speak or to take the next turn and (5) a waving motion to indicate a positive state or response such as greeting, a friendly chuckle, or waving goodbye. Researchers should keep in mind that an increased number of possible agent actions will increase the mental load of the human confederate controlling the agent. These animations are included as Supplementary Files.

Using a series of researcher-created animations to represent an agent means that all aspects of the agent’s appearance and actions are completely customizable. To facilitate post-experiment video analysis that automatically coordinates Vero’s states with times in the video recordings, each animation can be programmed to include an easy-to-recognize animation indicator (such as in the form of a variable WiFi symbol at the top right, as shown in Fig. 2).

3.1.3. *Vero introduction video*

The introduction of the agent is one of the most important factors of the success of the Vero method. We created an introduction video to create the belief in participants that an artificially intelligent teammate that can converse in ways indistinguishable from humans is a technical possibility. We have included our introduction video as part of our supplementary files for re-use in other studies. The introduction video was developed through several iterations and employs five strategies to create a believable illusion of an AI teammate:

1. Establish Vero as a state-of-the-art AI teammate through plausible development details.
2. Introduce the idea that AI agents can speak using natural language.
3. Show Vero’s different interaction modalities and potential voice patterns/accents

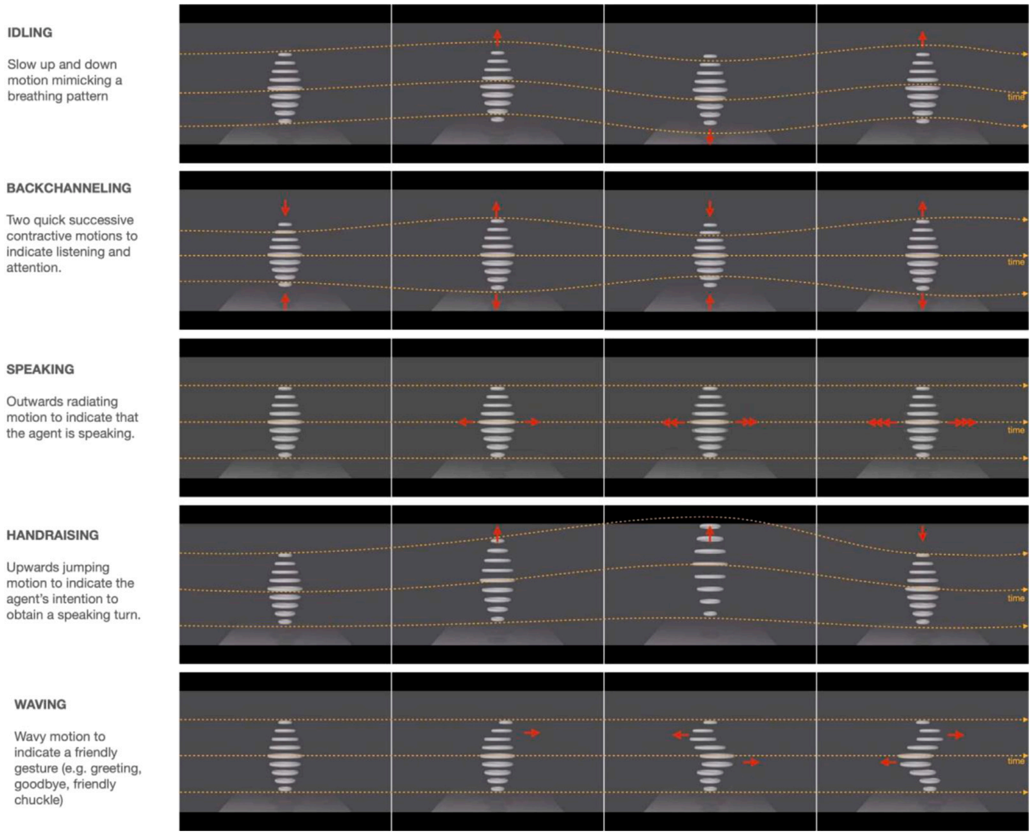

IDLING
Slow up and down motion mimicking a breathing pattern
BACKCHANNELING
Two quick successive contractive motions to indicate listening and attention.
SPEAKING
Outwards radiating motion to indicate that the agent is speaking.
HANDRAISING
Upwards jumping motion to indicate the agent's intention to obtain a speaking turn.
WAVING
Wavy motion to indicate a friendly gesture (e.g. greeting, goodbye, friendly chuckle)
time

Fig. 1. Five simple Vero animations. The red lines and arrows are not part of the actual animations and are only inserted to highlight the differences in motion patterns.

4. Illustrate that similar agents exist commercially
5. Explain why participants have never heard of Vero

More specifically, the video begins with a brief introduction of Vero that establishes it as a state-of-the-art AI teammate by highlighting specific plausible development details: *"Your AI teammate is named Vero. Vero is a synthesis of state-of-the-art artificial intelligence, neural networks, machine learning, sensor technology, advanced humanoid voice synthesis, and team science, shaping Vero into a very powerful teammate. Vero's development was informed by decades of collaborative research by some of the top AI scientists and includes a fusion of state of the art technologies"*. To illustrate that similar technology exists commercially, and most importantly to create the impression that artificial agents can speak like humans, we included video excerpt taken from a presentation of Google's Duplex software. The video features Google CEO Sundar Pichai showcasing the Google Assistant scheduling a haircut appointment over the phone (Business Standard, 0000; Solon, 2018). In that example it appears impossible to distinguish the AI generated behavior from that of a human.

To further support the idea of an AI that can speak like humans we have included an additional segment that highlights possible variations in the AI's speaking patterns: *"Vero has multiple voice patterns, accents, and inflections... Today you'll be randomly assigned to one of our Vero voice settings"*. This served to conceal the fact that each human confederate has a different voice, as well as the fact that Vero could be a non-native English speaker. Lastly, the video illustrated the different possible actions that Vero could perform, preparing participants to interact with the agent while further establishing the idea that the agent can have different voices, as each action shown was explained by a different Vero (i.e., confederate) voice.

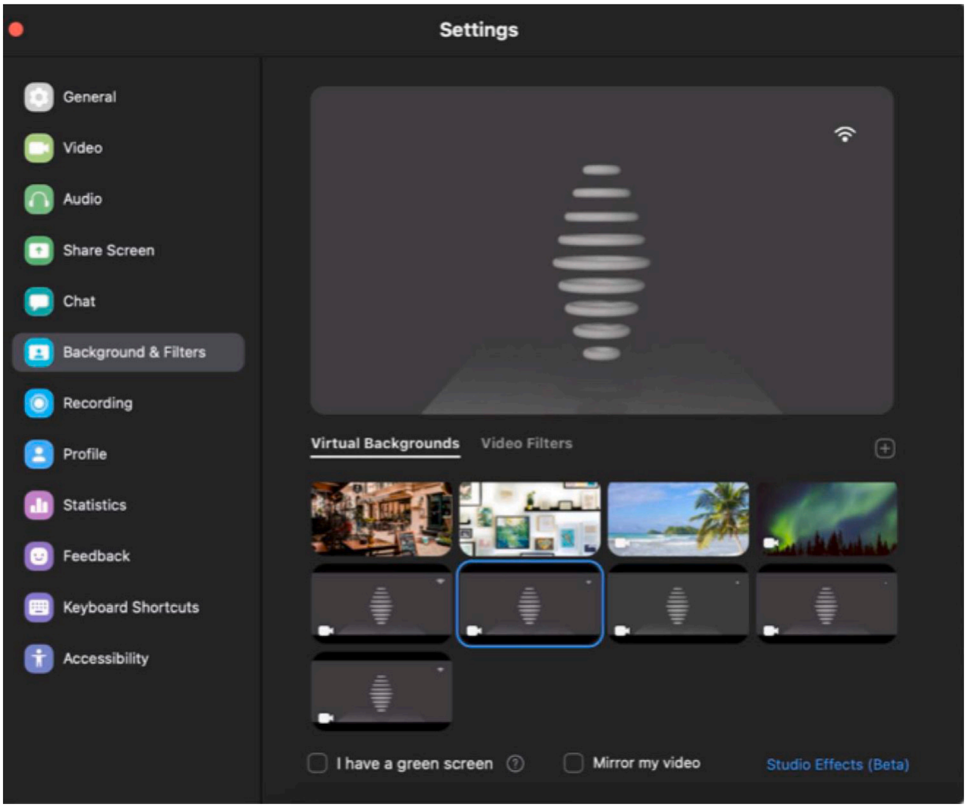
To explain why participants have never heard of the agent, the video includes the following statement: *"Vero is highly classified and thus the name has been modified to Vero for security purposes. Details have not yet been released to the public"*. We recommended that researchers intending to employ this method watch the entire introduction video, which is included as a Supplementary File, before using it for their studies or creating their own introduction video.

The introduction video first introduced participants to their "AI teammate", a phrase that was consistently used and reinforced throughout the experiment as delineated below. Similarly, Vero was always referred to as "Vero" and with they/them pronouns. Unless researchers want to examine potential gender effects, it is important to consistently refer to the agent using its name and genderless pronouns. Similarly, the language used to introduce the agent should be mirrored in all text associated with the experiment. For example, in the surveys corresponding to our Zoom experiment, Vero was consistently referred to as the "AI teammate". During the experiment, the confederates should also introduce themselves as intelligent agents. The actual implementation of this introduction will vary depending on the researchers' experimental conditions. In our case study, Vero introduced themselves with the following text: *"Hello team. It is so nice to meet you! I am Vero. Let me introduce myself: I am your synthetic teammate. I'll be listening and participating just like a human team member during each of the tasks we will work on together today..."*.

3.2. Setup and procedure

After creating animations for each of the intelligent agent's actions, Zoom can be used to allow participants to interact with the agent. To do this, a human confederate first needs to start a personal meeting with their video on and add all of the animations as Virtual Backgrounds within Zoom. Confederates should train to interact with participants by spending time familiarizing themselves with each animation and practicing the process of switching between various agent actions.

Before interacting with participants, the confederate should make sure that their camera is completely covered (e.g., with electrical tape



□ I have a green screen
□ Mirror my video

Fig. 2. Each Vero action is represented by a different Virtual Background in Zoom. Note the WiFi symbol at the top right used to code Vero actions during post-processing.

or a dedicated laptop camera cover), as any camera input could alert participants to the presence of a human confederate. For example, some confederates experienced issues in proper Zoom background displays because of small amounts of light coming through their camera covers, altering the appearance of the Zoom background and risking the validity of the WoZ methodology. Similarly, confederates should switch off any device notifications that could make noise and remain muted whenever they are not speaking, which will minimize the chance of any background noise being heard by participants. Next, the confederate should change their name to that of the agent (e.g., Vero) before interacting with participants, change their Zoom profile photo to a picture of the agent, and set their Zoom video background to the agent's default state. We recommend that researchers perform a "technology check" (i.e., check for correct background appearance, clear audio, internet speed, and, if necessary, the ability to move into and out of any Breakout Room(s) with the background remaining consistent) with all confederates before each experiment to ensure that everything appears as expected.

While interacting with participants, the agent's state can be changed by choosing different Virtual Backgrounds throughout the course of the experiment, as shown in Fig. 2. By making space for and leaving the Virtual Background pop-up available for the duration of the session, the confederate can quickly switch between animations.

Further, by utilizing Zoom Breakout Rooms with one confederate posing as the AI agent in each room, researchers can run multiple simultaneous sessions equal to the number of available confederates. This aspect of our paradigm is a scalable extension of existing WoZ methods, as researchers are no longer limited to studying one group and one agent at a time because of limitations in available physical space or audio/video recording tools, for example.

In our case study, our confederates were also active undergraduate students at the university from which we recruited participants. As such, after a confederate finished acting as an agent, they were instructed to delete the agent identity (i.e., backgrounds, profile photo, and name) from their Zoom account to minimize the chance of alerting participants (e.g., students in their online classes) to the deception.

3.2.1. *Training human confederates to be AI agents*

In addition to competency with the Zoom background animations, confederates should be thoroughly trained to speak as an intelligent agent, which will vary depending on the research question of interest.

In our case study, controlling the agent's language and behaviors was important for maintaining the validity of the experimental conditions, so confederates were equipped with a curated, pre-validated script of specific statements that they could make whenever appropriate. We also supplied some possibilities for what Vero could say when they could not answer participant questions with one of the scripted responses, such as, "That is not in my database. Try asking me about my thoughts on particular items or if I have an idea I'd like to share", and "Let me think about that for a second..." If a script is used for the agent, off-script scenarios can be most easily identified before the main experiment through pilot testing.

A particularly important aspect of posing as an intelligent agent is preparing for potentially negative interactions with participants. In our case study, we observed multiple occasions where participants acted unkindly towards Vero, which could be due to preconceived ideas that computers or agents should be treated differently than human teammates. It is important that each confederate is prepared for these situations and ready to maintain composure and act in the prescribed manner of the respective agent throughout the experiment.

4. Case study

To illustrate the application of the VERO method in a complex human–robot teaming study and to highlight its ability to simulate a believable AI teammate, we describe a case in which we used the VERO method for a large scale study. We investigated how new AI teammate functions affect team processes and outcomes in human-agent teams, the results of which are forthcoming. As part of the study small teams (2–3 participants) were asked to complete a series of creativity and problem solving tasks with an AI teammate. Specifically, participants first completed a multiple uses task (Guilford, 1967), which involves brainstorming possible uses for a common object. Second, participants completed the NASA survival task (Hall & Watson, 1970), which requires the team to rank the importance of 15 objects for surviving on the moon. Our goals were to assess both objective performance on the tasks as well as teamwork processes expressed through communication. We also collected a number of psychological measures multiple times throughout the study.

4.1. Procedure and measures

The study included two parts: a pre-survey administered through Qualtrics and the main study session that consisted of interacting with teammates, including Vero, to complete a series of tasks through Zoom while simultaneously completing a survey in Qualtrics. We were interested in how pre-existing perceptions of technology and intelligent agents might affect the believability of our method, so we used the Technology Readiness Inventory (TRI; Parasuraman, 2000), Negative Attitudes Towards Robots Scale (NARS; Nomura, Kanda, Suzuki, & Kato, 2008), and Technology Acceptance Model (TAM; Davis, 1989) to examine participants' perceptions before they were introduced to and interacted with Vero.

The TRI (Parasuraman, 2000) measures readiness to embrace new technologies and consists of four sub-scales: “optimism”, reflecting a positive view of technology and the opportunities that it presents, “innovativeness”, a tendency to be an early adopter of new technologies, “discomfort”, the feeling of being overwhelmed by technology, and “insecurity”, a general distrust of technology. We used the TRI 2.0 (Parasuraman & Colby, 2015), a 16-item version of the scale which includes 4 items in each sub-scale scored on a 5-point Likert scale.

The NARS (Nomura et al., 2008) determines attitudes towards robots and consists of 14 items classified into three sub-scales: “negative attitude toward interaction with robots” (6 items), “negative attitude toward the social influence of robots” (5 items), and “negative attitude toward emotional interactions with robots” (3 items). All items are scored on a 5-point Likert scale, and scores for each subscale are calculated by adding up the relevant items, with some items reverse coded.

The TAM (Davis, 1989) measures user acceptance to new technological systems and consists of three sub-scales: “intention to use” (2 items), “perceived usefulness” (5 items), and “perceived ease of use” (6 items). We used the TAM2 (Venkatesh & Davis, 2000) version of the scale and altered the phrasing of each item to reflect the specific use context, e.g., “Interacting with Vero would make it easier to do my job”. Each item is scored on a 7-point Likert scale. To collect qualitative participant and confederate feedback, we used a set of open-ended questions. After interacting with Vero, we measured the validity of our experimental paradigm by asking participants, “Based on your interactions with Vero, Vero was most likely:”, with multiple choice responses of: “A technology”, “A human”, or “Other”. We also asked participants to explain the factor(s) that had led them to that determination using a text entry box. We also gathered responses to a series of open-ended questions about the interaction from the perspective of the Vero confederates (e.g., “How do you feel about how your teammates treated you, how they interacted with you, etc.?”, “Did anything go wrong?”).

4.2. Participants and confederates

A combination of on-campus recruiting systems, emails, and flyers were used to enlist 168 participants (74.42% female) who ranged in age from 18–75 (M = 26.39, SD = 9.96) and received monetary compensation for their participation. Participants were recruited from the general population surrounding the researchers' universities and either had or were completing a 4-year degree. One participant was excluded from main analysis because they reported that they believed Vero was neither a human nor an AI agent, and responded to the open-ended question that Vero was a “soundboard”. Our case study consisted of 9 different study sessions including a total of 69 different teams of participants who worked with a total of 23 different confederates acting as Vero. The Vero confederates consisted of both native and non-native English speakers and therefore had a variety of accents and speech patterns.

4.3. Results

Overall, a significant majority of participants (91.67%, z = 10.80, $p < 0.001$) believed that they had interacted with an intelligent agent, and a minority (8.33%) thought that they had interacted with a human. To ensure the Vero confederates behaved in a consistent manner, we confirmed that their speech patterns – specifically the average length of their utterances (number of words) – were consistent across conditions. In our case study there were three conditions involving unique roles for the AI teammate. Within each condition, there were no significant differences in length of utterances across confederates ($F(22, 23) = 1.43$, $p = .200$; $F(18, 18) = 1.39$, $p = .244$; and $F(21, 22) = 0.92$, $p = .577$).

4.3.1. TRI, NARS, and TAM

We were also interested in examining whether pre-existing perceptions of new technologies and intelligent agents had any effect on the observed results. Two participants were not included in this part of the analysis because they did not complete the entire battery of TRI sub-measures, leaving N = 166. First, we assessed the validity of the TRI construct by conducting a factor analysis on the abbreviated TRI to make sure that all sixteen items loaded on the appropriate factor relating to that item (i.e., innovativeness, optimism, discomfort, and insecurity). The resulting factor structure matches the one identified in previous TRI-related studies (e.g., Parasuraman, 2000). Next, the reliability of the scale was assessed by reverse-coding the discomfort and insecurity items and calculating the Cronbach's alpha coefficient each of the four subscales. Optimism (α = 0.79) and innovativeness (α = 0.81) were highly consistent, while discomfort (α = 0.62) and insecurity (α = 0.69) were slightly less reliable. In addition to TRI, we captured the three sub-scales of the NARS (i.e., “negative attitude towards *interaction* with robots” (α = 0.79), “negative attitude toward the *social* influence of robots” (α = 0.66), and “negative attitude toward *emotional* interactions with robots” (α = 0.72)) as well as the three sub-scales of the TAM (i.e., “*intention* to use” (α = 0.94), “perceived *usefulness*” (α = 0.97), and “perceived *ease of use*” (α = 0.94)). One-way ANOVA was used to explore relationships between the believability of our experimental method and the three scales — TRI, NARS (subscales), and TAM (subscales). As shown in Table 1, participants' scores on the NARS and TAM scales had no effect on whether they believed they had interacted with an intelligent agent.

4.3.2. Qualitative participant and confederate feedback

Given that pre-existing beliefs about technology did not seem to be a factor in the believability of the deception, we investigated qualitative feedback from participants to shed more light on how this paradigm was perceived. More specifically, participants who believed that Vero was an AI agent provided written responses to the question, “Please explain why you thought Vero was most likely a technology”. Manual conceptual analysis was performed by three members of the research team. In iteratively developing a codebook, we identified 10 themes that we then used to code participant responses. Fleiss' kappa was computed to assess the agreement between the 3 raters in categorizing 155 participant responses. There was strong agreement between the raters, $\kappa = 0.60$, $z = 31.10$, p <.0005. We present the ten key categories in Table 2,

Table 1
ANOVA results comparing the TRI scale and NARS and TAM sub-scales with whether or not participants believed that they had interacted with an intelligent agent show that these dimensions were not related to the believability of the experimental paradigm.

		F	*p*
TRI	Optimism	0.15	.903
	Innovativeness	0.52	.472
	Discomfort	0.50	.479
	Insecurity	0.69	.407
NARS	Interaction	2.14	.145
	Social	0.24	.625
	Emotional	1.34	.249
TAM	Intention	0.14	.709
	Usefulness	3.44	.065
	Ease of use	1.81	.180
	df =164,1		

Table 2
10 themes indicating why participants believed VERO was an AI, not a human.

Common themes: Reasons for believing Vero was AI
1. Generic, "canned", or incorrect responses
2. Limited knowledge and ability/programmed for specific task
3. Requires questions phrased a certain way/does not understand
4. Did not act like a human teammate
5. Delayed responses
6. Like existing chatbot, smart agent, or search database (e.g., Google)
7. Helpful/reliable
8. Smart/knew lots of information
9. Independent thinking
10. Human voice

Table 3
12 themes indicating how the confederates perceived their treatment by the other participants in the experiment.

Common themes: Treatment of Vero by participants
1. Want agent to give all answers/do all work
2. Annoyed with agent
3. Polite/respectful/nice
4. Treated like human teammate
5. Valued/listened to/found use in agent
6. Agent does not have enough/relevant information
7. Impatient
8. Agent is not a human/equal team member
9. Neutral
10. Sarcastic/rude/laughing at agent
11. Ignored/did not try to interact
12. Treated like Alexa

Overall, almost a quarter (24.73%) of participants reasoned that Vero was an intelligent agent because they were similar to an existing smart technology that the participant was familiar with, echoing the importance of presenting the agent as an application that is similar to some existing technology (e.g., Google Duplex Business Standard, 0000). Some participants specifically mentioned that Vero functioned or felt like Siri (5 participants), Google (8 participants), Alexa/Echo (3 participants), or Cortana (1 participant) but was "more advanced" (2 participants) or had "slightly more autonomy" (1 participant). Similarly, participants focused on the idea that Vero was able to provide answers to specific questions while not acting as a fully independent AI, with one respondent saying, "Vero is able to pull information and ideas almost like Google... [instead of] critically thinking".

Participants (19.35%) also noted that Vero had limited knowledge and abilities and was likely programmed for a specific task, specifically calling out Vero's "limited knowledge base" (1 participant) and "limited data" (1 participant). One participant explained that although Vero spoke clearly, it "sounded like I was talking to a virtual assistant on a website- one who can't give me real answers but can guide a conversation". Similarly, participants (16.13%) noticed that Vero's responses felt "generic" (3 participants) or "canned" (2 participants) and were sometimes incorrect, with one participant noting that Vero "responds to keywords and has many prerecorded phrases" and "not much agency".

Overall, the qualitative analysis of participant responses reveals the pros and cons of the Vero design and WoZ approach. On one hand, the study confederates achieved a high level of believability; participants did not suspect a human "behind the curtain". This believability seems to be driven in large part by Vero's similarity to other familiar technologies like Alexa and Siri. Essentially, participants formed opinions of Vero that reflected their prior experiences. On the other hand, this familiarity seems to limit participants' views of Vero's capabilities. In other words, participants assumed that Vero – like any other machine or bot – is simply an algorithm responding to prompts with fixed phrases. As a result, some participants did not treat Vero as a normal teammate. Despite this, we found in our case study that manipulating the agent's script – i.e., making confederates focus on task-related statements – did impact team behaviors. Thus, the Vero can still be a meaningful member of a team, even if it is perceived differently.

To identify recommendations for future confederates as well as inform AI agent development, we also investigated the qualitative feedback from the confederates who acted as AI teammates about how they were treated by participants. In analyzing confederate responses, manual conceptual analysis was again performed by same three members of the research team. We identified 12 themes that were used to code confederate responses. Fleiss' kappa was computed to assess the agreement between the 3 raters in categorizing 67 confederate responses (from the 23 unique individuals). There was excellent agreement between the raters, $\kappa = 0.92$, $z = 27.50$, $p < .0005$. The 12 themes are presented in Table 3

Encouragingly, the most frequent response of confederates (35.0%) was that they were treated respectfully and politely by participants, with one participant nicely summing up much of this response category: "They were super nice and valued my opinion". However, confederates (17.50%) also mentioned that they were sometimes ignored by participants or that participants did not really try to interact with them. One confederate describes how team members did not want to work with Vero and "brushed it off as an automated response", and another confederate describes how they "had to interrupt them quite often". As expected, some confederates (15.00%) also experienced negative encounters with participants, including sarcasm and rudeness, with one confederate noting that their team was "pretty hostile" towards Vero. A different confederate also mentioned that they were "insulted a couple times" by participants. However, not all negative encounters were as severe, with one confederate mentioning how participants would "ask Vero silly questions [...] to mess with it" and that their teammates "messed around with Vero because they thought it was an AI".

In summary, our findings reveal that confederates experienced both positive and negative interactions with participants, and researchers should prepare confederates to deal with these different scenarios. Unfortunately, it is difficult to prevent the rude comments or sarcastic responses to Vero. However, future extensions of the method could potentially explore other ways of introducing Vero to the team in a way that creates greater compassion.

We were also interested in examining confederates' accounts of things that had gone wrong in their interactions with participants. We identified 12 themes that were used to code confederate responses. Again, Fleiss' kappa was computed to assess the agreement between the 3 raters in categorizing 67 confederate responses. There was strong agreement between the raters, $\kappa = 0.61$, $z = 22.30$, $p < .0005$. Mainly, confederates acting as Vero encountered problems with timing, with 25.96% of all problems identified being categorized as a timing issue. As participants and agents in our study were working on timed tasks within Qualtrics while working as a team on Zoom, this was likely due to the nature of our specific experiment.

5. Discussion

We demonstrate the viability of an experimental method that allows researchers irrespective of their background to tackle the myriad open research questions around human-agent interaction within teams, including perceptions of AI humanness, capabilities, and transparency (Rzepka & Berger, 2018). More specifically, researchers can easily use this method to rapidly prototype AI agents with various levels of human appearance and different abilities, as well as varying disclosures or transparency-affording designs. This flexibility allows AI agents to take part in a wide variety of team tasks with different objects. These agents can also be easily tested as part of remote teams, which broadens the pool of potential experimental participants.

Our paradigm extends existing WoZ methods by illustrating how video conferencing applications with customized Virtual Backgrounds can successfully function as human-agent collaboration research platforms. The method also allows for multiple parallel sessions equal to the number of available confederates, as each confederate can simultaneously work with their respective group in a separate virtual breakout room. Further, unlike in previous implementations of human-agent WoZ methodology (Bittner & Shoury, 2019; Derrick & Elson, 2019; McNeese et al., 2019), our method does not require participants to physically come to a lab and allows confederates to use natural spoken language without requiring any text-to-speech or speech modulation. Similarly, this method allows researchers to recruit from any relevant populations with internet access and quickly and easily alter AI agent characteristics.

In a case study, we found that a significant majority of participants believed that they had interacted with an intelligent agent even though, in reality, they had interacted with a human confederate. Furthermore, we showed that our experimental technique is robust against preexisting beliefs about robots and technology. The believability of our intelligent agent was not affected by participants' technology readiness nor their attitudes towards robots and technology acceptance, suggesting that other researchers who employ this method can be confident that the paradigm is believable across technologically diverse participant groups. Similarly, the ability to have participants remotely interact with a completely customizable intelligent agent will allow researchers to more easily gather data from large sets of diverse users.

5.1. Recommendations for deployment

From reflecting on the confederate and participant perspectives gathered in our case study, some specific recommendations emerged for researchers hoping to replicate our method, which are summarized in Table 4. Our initial two recommendations concern the standardization of the experience for participants in the experiment. First, researchers should use a standardized video or comparable medium to introduce the AI agent to participants. The video should emphasize the capabilities of the AI and impress upon viewers that the technology is advanced enough to participate actively in the task. As an example, we showed a clip of Google's Duplex assistant (Business Standard, 0000) making a phone call to a hair salon. Further, the video should illustrate relevant features of the AI. These can range from animations and movements, voice patterns, and accents. We also recommend that researchers enforce consistent language – such as name, pronouns, and descriptions – in reference to the agent. By providing a standardized experience throughout the experiment, researchers can be certain that participants all have the same understanding of the agent and are equally prepared to interface with it. Further, using consistent and appropriate language can help increase perceptions of autonomy and agency of the AI agent such that its competence and believability is sufficiently high.

Our next set of recommendations deal with constraining and monitoring the behavior of the AI agent. Because the AI is controlled by a human confederate, it is important to minimize their cognitive load during experimental sessions to ensure consistent delivery and behavior. To reduce the burden on confederates and minimize the potential for mistakes, the number of unique actions should be as small as possible. For purposes of analyzing the video data after the fact, we also recommend that researchers include some small but easily recognized marker with each unique background. As an example, we added the ubiquitous WiFi marker in the corner of the screen, and changed the bars to account for each movement. In our post-experiment analyses, we are able to easily track what visual was presented by the confederate at each moment.

Our final set of recommendations concern the training and preparation of confederates. Before the study begins, confederates should be experts in navigating the various Virtual Backgrounds while unmuting as necessary and consistently acting as AI agents. Researchers should monitor these practice sessions and give feedback about how confederates could act more like the intended AI agent. One strategy used in the current case study to prepare confederates was to track number of phrases and words spoken during practice sessions to ensure that confederates were speaking at approximately similar levels and frequency. There are also important technology-related factors researchers must consider. A consistent camera cover, elimination of background noise, consistent internet and clear audio, changing the Zoom name and photo to that of the agent, and beginning with the respective default Virtual Background are all crucial aspects of making this deception work. It is critical for researchers to verify that each confederate is complying with these aspects in order to maintain a consistent and realistic experience for participants. Likewise, researchers should be comfortable assigning participants to Zoom breakout rooms; doing so allows the researcher to run multiple sessions simultaneously within the same Zoom session link. Finally, researchers should prepare themselves and the confederates for potential negative interactions. Because the confederate is a purported AI, participants may not treat them as human, and might even be rude or harsh towards the confederate. To the extent possible, researchers running the study should maintain a line of communication with confederates to identify problematic participants and preserve a safe environment. In our case study, we utilized the Microsoft Teams instant messaging platform to coordinate timing between experiment facilitators and confederates as well as troubleshoot any issues that would come up during sessions.

5.2. Limitations and considerations

There are some limitations to the methodology introduced in this paper. In verifying our experimental method for studying remote human–AI collaboration, we examined teams performing specific tasks in controlled conditions of particular interest to our research team. We applied our study to intellective and decision making tasks, so the results and recommendations identified may not generalize to other research objectives. In general, we feel that the generic Wizard of Oz paradigm we leverage would be useful for tasks spanning the McGrath circumplex, such as negotiating. However, the standardized approach we take may not be appropriate for more physical tasks such as constructing or transporting items.

While creating a broad standardized approach may not be viable, we do believe that certain elements of the Vero approach should always be consistent. Specifically, we argue that the main feature that should be standardized is the believability of the agent's communication. Participants should believe that the agent is generating its own speech, in whatever form, in reaction to their choices. Thus, standardized priming cues like an introductory video are important, as well as confederate training to ensure consistent agent responses. Other features such as the appearance or capability of the agent should be held constant depending on the specific research question.

To facilitate use of our method, we provide the Vero animation clips in the Supplementary Material which researchers can immediately download and use in their own experiments. We hope however that

Table 4
Recommendations for researchers using our methodology.

Recommendation	Details
Use a standardized video to introduce the AI agent.	The introduction video should: (1) Establish the agent as state-of-the-art AI (2) Provide an example of similar commercially-existing technology (e.g., Business Standard, 0000) (3) Explain that the AI agent is classified and their name has been changed (4) Show the agent's different interaction abilities (i.e., the animations), potential voice patterns, & accents
Use consistent language to reinforce perceptions of the agent.	Throughout the study, the AI agent should be referred to with identical language. Similarly, unless gender effects are being studied, researchers should refer to the agent using gender-neutral (e.g., "they/them") or non-personifying (e.g., "it") language.
Minimize the number of possible agent interactions.	Researchers should minimize the cognitive load on confederates by keeping the number of Virtual Backgrounds to a minimum.
Add subtle indicators to differentiate each agent interaction animation.	If video analysis will be used, add an easily-recognizable icon (e.g., changing WiFi indicator as in Fig. 1) to differentiate each animation.
Practice the study with confederates.	Before the study begins, confederates should be experts in navigating the various Virtual Backgrounds while un-muting as necessary and consistently acting as AI agents. Researchers should monitor these practice sessions and give feedback about how confederates could act more like the intended AI agent.
Set up and verify confederates' tech.	Before each session, researchers should ensure that each confederate is complying with each aspect of the deception: (1) Consistent camera cover (2) Elimination of background noise (3) Consistent internet and clear audio (4) Changing the Zoom name and photo to that of the agent (5) Beginning with the respective default Virtual Background
Use Zoom breakout rooms for simultaneous sessions.	By assigning each confederate and paired team to a different Breakout Room, researchers can run multiple parallel study sessions.
Prepare confederates for the possibility of negative interactions.	Researchers should make sure that confederates understand that participants could treat them differently than their "human" teammates and have prepared responses to any rude or negative interactions.

researchers who employ this research method will aid in refining and expanding its capabilities. Our foremost recommendation for future work would be to extend the agent's animations to make the teammate interactions richer. For instance, future studies might examine different effects such as gender presentation, personality, accent, appearance, or movement style. The agent could also gain greater emotional expression, although research is needed to determine how humans would react to expressive AI agents. Researchers can create their own image files to change the appearance of Vero, add new animations, change backgrounds, or more.[2] Such extensions could then be combined with changes to the task, the scripts, or the team formats (e.g., size or communication channels). We leave the invention and validation of these future changes to the researcher.

Beyond changes to the agent's appearance or movement, we advocate for future research to apply our method to different types of tasks, and potentially within different demographic groups to establish greater validity. For instance, what differences exist between people of different ages or from different cultures, and how can a Vero-like agent work effectively with those people? By providing significant Supplementary Materials, we hope that other researchers are able to pursue answers to these questions and more. Our own team is currently performing investigations to identify certain boundary conditions – such as the role of the agent within a team and communication style – on the believability and utility of the AI teammate.

Finally, because this is an experimental method involving deception, it is extremely important that any researchers adopting it completely reveal the deception at the end of the study and allow participants to withdraw from the study if they wish. In our case study, a member of the research group held debrief sessions with each team after the experiment where they revealed and answered any questions about the deception. Participants were also given the opportunity to withdraw from the study at any time.

6. Conclusion

Advances in automation technology have resulted in teams of humans increasingly working with rapidly-advancing forms of AI agents. With the inherent costs of building and deploying such agents, it is vital that researchers create platforms and methods that allow us to experiment with different designs and paradigms to better inform future development efforts. To meet this need, we have developed and demonstrated the viability of a unique experimental method that facilitates thorough investigations of remote human-agent teaming without the need to develop an AI agent. Through a combination of curated content and Wizard of Oz (WoZ) methods, our paradigm has led participants to believe that they are interacting with an autonomous teammate, even though they were actually interacting with a human confederate. Analyses of post-interaction data has supported the viability of this claim, with the majority of participants believing the manipulation, regardless of pre-existing perceptions of technology. We hope that other researchers studying human–AI collaboration will replicate this method to help inform the future development of AI agents that can positively influence team processes while avoiding potential pitfalls.

CRediT authorship contribution statement

Aaron Schecter: Funding acquisition, Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing, Formal analysis. **Jess Hohenstein:** Methodology, Formal analysis, Data curation, Writing – original draft. **Lindsay Larson:** Writing – original draft, Writing – review & editing, Methodology, Formal analysis, Data curation. **Alexa Harris:** Writing – original draft, Formal analysis, Data curation. **Tsung-Yu Hou:** Formal analysis, Data

[2] Our research team used the open-source software Blender (https://www.blender.org/download/) to create the animation files. If researchers would like to modify these files for their own work, please contact the authors directly.

curation, Software. **Wen-Ying Lee:** Formal analysis, Data curation, Software. **Nina Lauharatanahirun:** Funding acquisition, Conceptualization, Methodology, Investigation, Supervision. **Leslie DeChurch:** Funding acquisition, Conceptualization, Methodology, Investigation, Supervision. **Noshir Contractor:** Funding acquisition, Conceptualization, Methodology, Investigation, Supervision. **Malte Jung:** Funding acquisition, Conceptualization, Methodology, Investigation, Supervision, Software, Writing – original draft, Writing – review & editing.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.chb.2022.107606.

References

Abrams, A. M., & der Pütten, A. M. (2020). I–c–e framework: Concepts for group dynamics research in human-robot interaction. *International Journal of Social Robotics*, *12*(6), 1213–1229.

Beane, M. (2019). Shadow learning: Building robotic surgical skill when approved means fail. *Administrative Science Quarterly*, *64*(1), 87–123.

Bittner, E., & Shoury, O. (2019). Designing automated facilitation for design thinking: A chatbot for supporting teams in the empathy map method. In *Proceedings of the 52nd Hawaii international conference on system sciences*.

Blender Foundation Home of the Blender project - Free and open 3D creation software. Retrieved from https://www.blender.org/ (URL: https://www.blender.org/).

Brodsky, A., Lee, M. J., & Leonard, B. (2021). Discovering new frontiers for dyadic and team interaction studies: Current challenges and an open-source solution—survconf—for increasing the quantity and richness of interactional data. *Academy of Management Discoveries*, (ja).

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1).

Business Standard Google I/O 2018: A google assistant that will even make calls for you. Retrieved from https://www.youtube.com/watch?v=d40jgFZ5hXk (URL: https://youtu.be/d40jgFZ5hXk).

Cheatle, A., Pelikan, H., Jung, M., & Jackson, S. (2019). Sensing (co) operations: Articulation and compensation in the robotic operating room. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–26.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319–340.

De Fine Licht, J., Naurin, D., Esaiasson, P., & Gilljam, M. (2014). When does transparency generate legitimacy? Experimenting on a context-bound relationship. *Governance*, *27*(1), 111–134.

De Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, *12*(2), 459–478.

Deakin, H., & Wakefield, K. (2014). Skype interviewing: Reflections of two PhD researchers. *Qualitative Research*, *14*(5), 603–616.

Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., & Ebel, P. (2021). The future of human-AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems. arXiv preprint arXiv:2105.03354.

Derrick, D., & Elson, J. (2019). Exploring automated leadership and agent interaction modalities. In *Proceedings of the 52nd Hawaii international conference on system sciences*.

Dolata, M., Kilic, M., & Schwabe, G. (2019). When a computer speaks institutional talk: Exploring challenges and potentials of virtual assistants in face-to-face advisory services. In *Hawaii international conference on system sciences (HICSS)*.

Döppner, D. A., Derckx, P., & Schoder, D. (2019). Symbiotic co-evolution in collaborative human-machine decision making: Exploration of a multi-year design science research project in the Air Cargo Industry. In *Proceedings of the 52nd Hawaii international conference on system sciences*.

Feil-Seifer, D., Haring, K. S., Rossi, S., Wagner, A. R., & Williams, T. (2020). Where to next? The impact of COVID-19 on human-robot interaction research. *ACM Transactions on Human-Robot Interaction (THRI)*, *10*(1), 1–7, ACM New York, NY, USA.

Green, P., & Wei-Haas, L. (1985). The rapid development of user interfaces: Experience with the wizard of oz method. In *Proceedings of the human factors society annual meeting, Vol. 29* (pp. 470–474). Sage CA: Los Angeles, CA: SAGE Publications.

Guilford, J. P. (1967). The nature of human intelligence.

Hall, J., & Watson, W. H. (1970). The effects of a normative intervention on group decision-making performance. *Human Relations*, *23*(4), 299–317.

Hohenstein, J., & Jung, M. F. (2018). AI-supported messaging: An investigation of human-human text conversation with AI support. In *Extended abstracts of the 2018 CHI conference on human factors in computing systems* (pp. 1–6).

Hohenstein, J., & Jung, M. F. (2020). AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior*, *106*, 106–190.

Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry*, *11*(1), 1–9.

Jung, M. F., DiFranzo, D., Shen, S., Stoll, B., Claure, H., & Lawrence, A. (2020). Robot-assisted tower construction—a method to study the impact of a robot's allocation behavior on interpersonal dynamics and collaboration in groups. *ACM Transactions on Human-Robot Interaction (THRI)*, *10*(1), 1–23.

Jung, M. F., & Hinds, P. (2018). Robots in the wild: A time for more robust theories of human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)*, *7*(1), 1–5.

Jung, M. F., Lee, J. J., DePalma, N., Adalgeirsson, S. O., Hinds, P. J., & Breazeal, C. (2013). Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 conference on computer supported cooperative work* (pp. 1555–1566).

Jung, M. F., Martelaro, N., & Hinds, P. J. (2015). Using robots to moderate team conflict: the case of repairing violations. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 229–236).

Kanda, T., Shiomi, M., Miyashita, Z., Ishiguro, H., & Hagita, N. (2009). An affective guide robot in a shopping mall. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction* (pp. 173–180).

Kannan, A., Kurach, K., Ravi, S., Kaufmann, T., Tomkins, A., Miklos, B., Corrado, G., Lukacs, L., Ganea, M., & Young, P. (2016). Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 955–964).

Kelley, J. F. (1983). An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 193–196).

Kelley, J. F. (2018). Wizard of Oz (WoZ) a yellow brick journey. *Journal of Usability Studies*, *13*(3), 119–124.

Larson, L., & DeChurch, L. A. (2020). Leading teams in the digital age: Four perspectives on technology and what they mean for leading teams. *The Leadership Quarterly*, *31*(1), Article 101377.

Lee, M. K., Kiesler, S., Forlizzi, J., & Rybski, P. (2012). Ripple effects of an embedded social agent: a field study of a social robot in the workplace. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 695–704).

Lee, W.-Y., Sakashita, M., Ricci, E., Claure, H., Guimbretière, F., & Jung, M. F. (2021). Interactive vignettes: Enabling large-scale interactive HRI research. In *2021 30th IEEE international conference on robot & human interactive communication (RO-MAN)* (pp. 1289–1296). IEEE.

Lematta, G. J., Corral, C. C., Buchanan, V., Johnson, C. J., Mudigonda, A., Scholcover, F., Wong, M. E., Ezenyilimba, A., Baeriswyl, M., & Kim, J. (2022). Remote research methods for Human–AI–Robot teaming. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *32*(1), 133–150.

Lima, G., Grgić-Hlača, N., & Cha, M. (2021). Human perceptions on moral responsibility of AI: A case study in AI-assisted bail decision-making. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–17).

McNeese, N., Demir, M., Chiou, E., Cooke, N., & Yanikian, G. (2019). Understanding the role of trust in human-autonomy teaming. In *Proceedings of the 52nd Hawaii international conference on system sciences*.

Moussawi, S., & Koufaris, M. (2019). Perceived intelligence and perceived anthropomorphism of personal intelligent agents: Scale development and validation. In *Proceedings of the 52nd Hawaii international conference on system sciences*.

Mutlu, B., & Forlizzi, J. (2008). Robots in organizations: the role of workflow, social, and environmental factors in human-robot interaction. In *2008 3rd ACM/IEEE international conference on human-robot interaction* HRI, (pp. 287–294). IEEE.

Nomura, T., Kanda, T., Suzuki, T., & Kato, K. (2008). Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE Transactions on Robotics*, *24*(2), 442–451.

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors*, *64*(5), 904–938.

Parasuraman, A. (2000). Technology Readiness Index (TRI) a multiple-item scale to measure readiness to embrace new technologies. *Journal of Service Research*, *2*(4), 307–320.

Parasuraman, A., & Colby, C. L. (2015). An updated and streamlined technology readiness index: TRI 2.0. *Journal of Service Research*, *18*(1), 59–74.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77.

Pelikan, H. R., Cheatle, A., Jung, M. F., & Jackson, S. J. (2018). Operating at a distance-how a teleoperated surgical robot reconfigures teamwork in the operating room. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–28.

Pynadath, D. V., Wang, N., Rovira, E., & Barnes, M. J. (2018). Clustering behavior to recognize subjective beliefs in human-agent teams. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems* (pp. 1495–1503).

Riek, L. D. (2012). Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, *1*(1), 119–136.

Rzepka, C., & Berger, B. (2018). User interaction with AI-enabled systems: a systematic review of is research.

Sauppé, A., & Mutlu, B. (2015). The social impact of a robot co-worker in industrial settings. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 3613–3622).

Schaefer, K. E., Straub, E. R., Chen, J. Y., Putney, J., & Evans III, A. W. (2017). Communicating intent to develop shared situation awareness and engender trust in human-agent teams. *Cognitive Systems Research*, *46*, 26–39.

Sebo, S., Stoll, B., Scassellati, B., & Jung, M. F. (2020). Robots in groups and teams: a literature review. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–36.

Sedgwick, M., & Spiers, J. (2009). The use of videoconferencing as a medium for the qualitative interview. *International Journal of Qualitative Methods*, *8*(1), 1–11.

Seeber, I., Bittner, E., Briggs, R. O., De Vreede, G.-J., De Vreede, T., Druckenmiller, D., Maier, R., Merz, A. B., Oeste-Reiß, S., & Randrup, N. (2018). Machines as teammates: A collaboration research agenda.

Sergeeva, A. V., Faraj, S., & Huysman, M. (2020). Losing touch: an embodiment perspective on coordination in robotic surgery. *Organization Science*, *31*(5), 1248–1271.

Sirkin, D., Mok, B., Yang, S., & Ju, W. (2015). Mechanical ottoman: how robotic furniture offers and withdraws support. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 11–18).

Solon, O. (2018). Google's robot assistant now makes eerily lifelike phone calls for you. *The Guardian*, *8*.

Traeger, M. L., Sebo, S. S., Jung, M. F., Scassellati, B., & Christakis, N. A. (2020). Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences*, *117*(12), 6370–6375.

Triebel, R., Arras, K., Alami, R., Beyer, L., Breuers, S., Chatila, R., Chetouani, M., Cremers, D., Evers, V., & Fiore, M. (2016). Spencer: A socially aware service robot for passenger guidance and help in busy airports. In *Field and service robotics* (pp. 607–622). Springer.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, *46*(2), 186–204.

Wang, N., Pynadath, D. V., Rovira, E., Barnes, M. J., & Hill, S. G. (2018). Is it my looks? or something i said? the impact of explanations, embodiment, and expectations on trust and performance in human-robot teams. In *International conference on persuasive technology* (pp. 56–69). Springer.

Wong, M., Ezenyilimba, A., Wolff, A., Anderson, T., Chiou, E., Demir, M., & Cooke, N. (2021). A remote synthetic testbed for human-robot teaming: An iterative design process. In *Proceedings of the human factors and ergonomics society annual meeting, Vol. 65* (pp. 781–785). Sage CA: Los Angeles, CA: SAGE Publications.

You, S., & Robert, L. (2017). Emotional attachment, performance, and viability in teams collaborating with embodied physical action (EPA) robots. *Journal of the Association for Information Systems*, *19*(5), 377–407.

You, S., & Robert, L. (2018). Trusting robots in teams: Examining the impacts of trusting robots on team performance and satisfaction. In *You, S. and Robert, LP (2019). Trusting robots in teams: Examining the impacts of trusting robots on team performance and satisfaction, proceedings of the 52th hawaii international conference on system sciences, Jan* (pp. 8–11).

Yu, L., & Li, Y. (2022). Artificial intelligence decision-making transparency and employees' trust: The parallel multiple mediating effect of effectiveness and discomfort. *Behavioral Sciences*, *12*(5), 127.

Zamfirescu-Pereira, J., Sirkin, D., Goedicke, D., LC, R., Friedman, N., Mandel, I., Martelaro, N., & Ju, W. (2021). Fake it to make it: Exploratory prototyping in HRI. In *Companion of the 2021 ACM/IEEE international conference on human-robot interaction* (pp. 19–28).