Sushi with Einstein: Enhancing Hybrid Live Events with LLM-Based Virtual Humans

Alon Shoa
The Advanced Reality Lab, Reichman University
Herzliya, Israel
shoa.alon@gmail.com

Ramon Oliva
Event Lab, Department of Clinical Psychology and Psychobiology, University of Barcelona
Barcelona, Spain
ramon.oliva.martinez@gmail.com

Mel Slater
Event Lab, Department of Clinical Psychology and Psychobiology, University of Barcelona
Event Lab, Institute of Neurosciences, University of Barcelona
Barcelona, Spain
melslater@gmail.com

Doron Friedman
The Advanced Reality Lab, School of Communications, Reichman University
Herzliya, Israel
doronf@runi.ac.il

Figure 1: A screenshot from the multi-user VR session: Albert Einstein.

IVA '23, September 19–22, 2023, Würzburg, Germany

ACM ISBN 978-1-4503-9994-4/23/09.
https://doi.org/10.1145/3570945.3607317

ABSTRACT

It is becoming increasingly easier to set up multi-user virtual reality sessions, and these can become viable alternatives to video conference in events such as international conferences. Moreover, it is possible to enhance such events with automated virtual humans, who may participate in the discussion. This paper presents the behind-the-scenes work of a panel session titled "Is virtual reality genuine reality?", which was held during a physical symposium, "XR for the people," in June 2022. The panel featured a

virtual Albert Einstein, based on a large language model (LLM), as a panelist, alongside three international experts having a live conference panel discussion. The VR discussion was broadcast live on stage, and a moderator was able to communicate with both the live audience, the virtual world participants, and the virtual agent (Einstein). We provide lessons learned from the implementation and from the live production, and discuss the potential and pitfalls of using LLM-based virtual humans for multi-user VR in live hybrid events.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; **User interface management systems**; *Empirical studies in collaborative and social computing*; • **Computing methodologies** → *Natural language generation*; • **Social and professional topics** → *Codes of ethics.*

KEYWORDS

VR, AI, Persona Reconstruction, GPT3

ACM Reference Format:
Alon Shoa, Ramon Oliva, Mel Slater, and Doron Friedman. 2023. Sushi with Einstein: Enhancing Hybrid Live Events with LLM-Based Virtual Humans. In *ACM International Conference on Intelligent Virtual Agents (IVA '23), September 19–22, 2023, Würzburg, Germany*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3570945.3607317

1 INTRODUCTION

Virtual reality (VR) is typically experienced in isolation from the real world. Here we provide an example of an event that was both experienced by participants in VR but was also a public physical event, and one attended also by a very illustrious guest. In June 2022, we had the opportunity to produce a live panel discussion, held in a multi-user VR environment, and live broadcast in front of an audience during a physical symposium. The panel, titled "Is Virtual reality a genuine reality?", featured a virtual Albert Einstein as a panelist alongside three real panelists: Professors David Chalmers, Mel Slater, and Doron Friedman, attending from three different continents in the same shared VR. The VR discussion was broadcast live on a screen, and a moderator on stage communicated with both the live audience as well as with the participants inside the VR. The virtual Albert Einstein took part in the discussion based on a voice interface and a large language model (LLM).

2 BACKGROUND

Multi-user VR environments have been around for a long time, but they are rarely used for hybrid events. Steed et al. [20] describe an asymmetric framework for social interactions, such that a physical space is enhanced to include remote visitors. Their "Beaming" platform also included the opportunity to replace participants by software controlled avatars, referred to as proxies [9, 10, 12, 15]. Our event has similar elements, and also some differences, mainly: i) we have tried to integrate a multi-user VR space with a physical space, rather than having just a physical space in Beaming, ii) unlike the proxy, the virtual Einstein is a fully autonomous character that does not intend to represent any of the real participants, and iii) the virtual autonomous guest is based on contemporary natural language processing (NLP) methods, which allows it to take part in a conversation. The event described here can also be considered a follow up of [17]; here we have added an "AI" based virtual participant, and also introduced a live audience.

Virtual humans portraying specific historic figures have been around for some time (e.g., [16]). A "live performance" by the artist Tupac in 2012 (using standard projection and visual effects), fifteen years after he was shot dead, captured the imagination of the wide public. Traum et al. [21] describe capturing a holocaust survivor, with automatic dialogue capabilities, based on speech and natural language understanding and playback of the most appropriate prerecorded answers.

Large language models have recently revolutionized NLP (e.g., [4, 7]). While such models are not specifically trained for dialogue, their capability for next word prediction can be harnessed towards dialogue. In the past this has required model fine tuning [5]. However, with the appearance of GPT-3 it is not clear whether models that have processed such huge amounts of text, including dialogue texts, actually require further training ("fine tuning", e.g., [14]) for dialogue; the current practice suggests that so-called 'prompt engineering' [18] is sufficient. Our main goal was to explore whether and to what extent a virtual human based on a contemporary LLM can take part in a multi-user VR conversation, seamlessly and naturally. The secondary goal was to explore whether this new possibility of adding virtual famous persona can be part of a hybrid live event.

3 ETHICAL CONSIDERATIONS

Recreating historical figures, especially in live public events, raises some social, legal, and ethical questions. It could raise concerns regarding the cynical use of public figures, as well as contributing to the so-called negative trend of "deep fake" [22].

It is not unlikely that LLMs may misrepresent the heritage of the persona being reconstructed. There is a question of the extent to which the memory of a public figure belongs to everyone, as it has become part of local or global culture, versus the extent to which the memory is private; this may be especially problematic in the context of figures who only recently passed away, and whose relatives may still be alive [3].

On the other hand, it can be argued that various forms of duplication and 'simulacra' are not new [1, 2], and "artificial intelligence"-based reconstructions do not pose any qualitative differences or new concerns. In this light, perhaps talking with a virtual replica of a historical figure is not different from viewing their portraits, reading their books, listening to their music – all of these can be reconstructed and transformed.

In the context of a live show in a virtual world with a panel of humans having a conversation with a VR persona of a historical figure, such as Albert Einstein, the following considerations, as outlined by [19], should be taken into account:

- **Privacy and Data Issues**: Ensure data privacy and prevent misuse.
- **Content-Induced Risk**: Respect and accurately represent Einstein's legacy.
- **Trust and Transparency**: Clearly communicate the nature of the VR persona.

- **Regulation and Responsibility**: Adhere to ethical guidelines and regulations.
- **Public Perception and Social Impact**: Consider the impact on public perception.
- **Potential Misuse**: Take measures to prevent misuse of the VR persona.

To ensure the ethical conduct of the live show, the following actions were taken in this project: we made it clear that the audience is aware of the nature of the VR persona as an LLM-based virtual agent. We have attempted a respectful and accurate visual representation of Albert Einstein. We implemented human-operated oversight of the generated content for responsible outputs (see below) in order to make sure there were no mis-representations.

4 METHOD

4.1 VRUnited

The panel session was held in a multi-user VR platform called VRUnited, developed by the EventLab, designed to enable participants to interact with each other using look-alike avatars and embodiment based on hand tracking with real-time inverse kinematics (IK) [17]. This platform enabled the conversation to take place in different locations around the globe, while live broadcasting it to the audience on screen. The system supports multiple virtual cameras recording from multiple sources, by connecting with invisible characters that camera operators can control, either in VR or using a PC.

4.2 MILO

MILO is a virtual agent framework designed to enable easy integration of "AI-controlled" virtual humans with dialogue capabilities to XR client applications. It includes a seamless integration of speech-to-text, LLM-based dialogue models, and text-to-speech. It is designed to integrate with animated virtual avatars in VR, and serves as the basis of several ongoing research projects. For the study reported here MILO was integrated with an Einstein look-alike avatar in the VRUnited system.

The text-to-speech (TTS) and speech-to-text (STT) used Google API. We set the system to have a German accent for Einstein, even though the conversation took place in English. The result was a bit difficult to understand at some points, but contributed to realism.

Natural language dialogue was based on the most recent version of GPT available at that time (GPT-3 [4], version DaVinci-002). A human operator manually monitored the conversation and could override the content in real-time, before allowing it to be broadcast to the audience; this manual censorship was a requirement of GPT-3 terms of service (ToS) at the time of operation. The ToS also require that all users (or audience) be made aware that this was not a real person, which was the case in the live event. Beyond the ToS requirements, it is useful to allow the operator to go through the script in real time for additional reasons, e.g., the operator could correct STT errors. In some applications it may be desired for a human operator to intervene in what the virtual character would say – not only remove offensive and inappropriate content but also improve the output; we did not allow this in the live event. We also used an automatic content-filtering provided by OpenAI's API to reduce the probability of offensive content being generated.

Prompt Design

Below is the full transcription of the panel "Is Virtual reality a genuine reality".
participants:
Prof. David Chalmers,
Prof. Mel Slater,
Prof. Doron Friedman,
And Prof. Albert Einstein as a special guest.

Conversation:

Chalmers: "One of the main points in my book is that virtual reality is genuine reality, people can lead meaningful lives even if they spend most of their life inside VR."
Slater: "However, my suggestion is that VR systems can be sorted such that system A is stronger than system B if system B can be simulated by system A but not the other way around."
Friedman: "If we assume that we all live in Plato's cave, then maybe VR is actually the way out of the cave..."
Einstein:

Figure 2: The initial prompt selected for the live event.

The prompt design (Figure 2) includes an initial description of the panel, with its participants (including Albert Einstein), followed by a few lines of discussion transcript (fake). As the real conversation begins, the utterances of the participants are concatenated to the prompt. If the length of the prompt exceeds the possible space (4096 tokens, where statistically a token is approximately 3/4 of a word, or 100 tokens represent approximately 75 words), the prompt is cut such that it includes the description of the panel and the most recent part of the conversation to fill the token quota.

A live transcription of the audio was based on Google streaming API. The operator could select from a list of models (LLMs), which could be different in terms of their hyper-parameters, fine-tuning, prompts, or more. In the event described here, we opted for using a generic pretrained LLM, but we had to carefully explore the precise prompt. With the specific model (DaVinci-002) we learned that if the prompts were too long the likelihood of the conversation derailing completely out of context increased, so we did not "use" the 4K token buffer in full. Also, the model was sensitive to tokens such as newline and ':'.

4.3 VRUnited-MILO Integration

The integration of MILO and VRUnited was based on Unity client scripts that initiate the conversation and send the audio to the MILO API via RTP/UDP. In order to simplify multi-user communication the applications was set such that participants had to press a button before they started to speak; testing under conditions of natural conversation are left for future work. The input audio was streamed to the STT component, and the transcription was presented on the operator dashboard. The operator could generate responses based on the given text, or modify it inside the dashboard. Figure 3 includes a schematic diagram.

MILO is a server component that deals with dialogue. In addition, there is a Unity component, which will refer to as Einstein client – this is a collection of scripts that take care of: i) Einstein animation (body language and lip sync), and ii) communication with MILO. We have designed and implemented a very simple API between the Einstein client (Unity, VRUnited) and MILO (server), including four types of messages:

(1) Start conversation – The Einstein client initiates a connection request to the MILO server, and a 'start conversation'

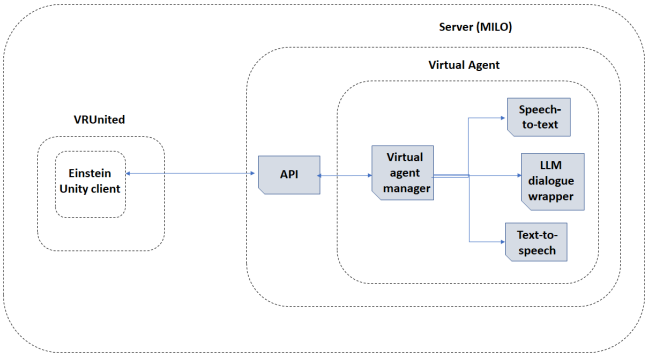

Figure 3: The high-level design diagram for MILO, including the interface with the VR client.

message is received from MILO when it is successfully connected to the VRUnited restaurant.

(2) End Conversation – the Einstein client sends this message to MILO in order to disconnect.
(3) Stream Audio – stream RTP (UDP) packets from the Einstein client (Unity/VRUnited) to the MILO server into a particular port. In order to simplify working in a multi-user setup the Einstein client changes the seed of the audio stream to indicate the identity of the speaker.
(4) Receive Ready – when MILO has a response generated ready and evaluated by the moderator, MILO sends a signal to the Einstein client and it retrieves the audio as a file from a specified location.

An interesting challenge for an autonomous system to take part in a multi-party discussion is understanding who the speaker is, for every utterance. We included two mechanisms for speaker identification: one in the VR side, described above. Similar functionality, assuming semi-automatic control, was also included in the server side (MILO): the operator UI included buttons with the names of the panel participants, and the operator was expected to press these, in real time, to denote the current speaker.

The Einstein model was developed from a head shot using Character Creator (Figure 1). The body language was based on a simple controller with two states: speaking and listening, with an animation loop for each. A snapshot of the event appears in Figure 4 and a video is available in[1].

4.4 Live Production Setup

The following roles were assigned during the production process:

- Panelists: the session included three panelists in three different continents. Only one of them was in the same physical location of the event.
- Moderator: on stage, served as the interface between the panelists who are immersed in VR and the audience outside VR.
- Physical videographer: in the auditorium, off stage, filming the audience and the stage.

[1]Companion video: https://www.youtube.com/watch?v=9bGUY9T4XVI

Figure 4: A view of the moderator and the projection of the VR on stage, as seen from the virtual director's position.

- Virtual director: in the auditorium next to the stage, controlling the application on a desktop interface (non-VR) with a laptop.
- AI operator: in the auditorium, off stage, controlling the MILO operator dashboard.
- Audio and setup technicians.
- Production manager: in touch with all participants through instant messaging for coordination.

Audio feedback was one of the main challenges. For the event we planned multiple input and output sources to be on stage – the local panelist and moderator were required to both talk and hear the session, and the audience in the auditorium was supposed to hear both the physical session as well as the VR session clearly. Eventually we resorted to having only the moderator on stage, in order to have only one audio input source.

Documenting the live session for later offline viewing raises additional challenges and opportunities. There are multiple sources of video: i) footage from the virtual director (desktop), ii) footage from the participants (HMD recording), and iii) footage from physical cameras. The virtual director moved the camera around during the session to establish various shots, and also instructed the participants on how to capture video segments and still images. However, participants controlled the virtual camera with their heads, inside VR, which resulted in unstable feed. This leaves the editor of the summary video with the dilemma on how to merge stationary shots with unstable shots taken with HMDs.

5 RESULTS

The event took place as part of a physical conference, in a university auditorium, with an audience of approximately 100 participants. The moderator gave an introduction about the panel discussion, which was focused on David Chalmers' new book, Reality+. At the same time, the panel participants joined the multi-user VR and waited for the moderator to kick-start the discussion, and the

VR was projected on stage, from the point of view of the virtual director. The virtual director, AI operator, and audio technician were all based off-stage in the main auditorium. A physical videographer was placed at the far corner of the auditorium.

The discussion took place among the VR participants, with occasional interventions from the moderator. Virtual Einstein was listening and generating potential text comments after every utterance, but the AI operator only sent these comments when Einstein was explicitly addressed by one of the participants or the moderator. The text that was automatically generated online was appropriate to the context of the discussion, and the summary statement can be considered inspiring (see companion video). Interestingly, the text included some German words infused into English.

The speech-to-text received input with three different accents (Israeli, English, and Australian), and the VR audio mixing performed quite well, but clearly not 100%. From observation, we saw that there were different types of speech recognition errors like missing words (deletion) and miss detection (substitutions). These errors can significantly degrade the performance, as compared to text only, which doesn't have these types of errors.

The moderator was connected to a desktop version of VRUnited. This was generally muted, unless when the moderator talked to the VR panelists – for example, taking questions from the audience and repeating them to the panelists. Following the session we collected all video footage and an editor prepared an edited version (see companion video).

6 DISCUSSION

We have addressed two main challenges. The first is how to produce a hybrid conference where the audience is in a physical auditorium and the participants are all in remote parts of the world. To the audience, the panelists seemed to be all in a restaurant talking to each other at one table, whereas in practice they were all based in different locations. This may be considered richer and more visually appealing than projecting video conference sessions on stage, and, of course, it is possible to develop this into arbitrarily complex virtual scenarios in future events. Since the virtual cinematographer used a desktop, the point of view was steady; we do not recommend projecting live footage from VR participants, since every small head movement is accentuated on the large projection for the stationary audience, which is confusing and might induce simulation sickness. One of the main challenges in such hybrid events is handling audio, such that each virtual or physical participants receives exactly one audio feed from all audio sources.

The main novel challenge was introducing an AI-controlled panelist. Thanks to recent progress in DNNs and LLMs, resulting in improved voice understanding and, especially, in dialogue capabilities, the conversation was mostly flowing and meaningful. A major next challenge will be to automate the role played by the human operator. In this case, the human operator was required by OpenAI ToS, but in other cases an automated agent that takes part in multi-party conversations would be highly desired; this requires the ability to understand turn taking and proactively blending in a multi-party conversation. There has been some work on automatic turn taking in virtual agents (e.g., [6]), but mostly in the context of dyadic conversations. In multi-party sessions this would require automatically understanding who the speaker is, and automatically deciding when to intervene (and when not to intervene) in the discussion.

Another challenge is automated cinematography – automatically or semi-automatically creating a video experience from events taking place in virtual environments. In fact this involves two types of challenges: i) real-time editing (what to display during the live event on the main screen), and ii) offline editing of a video summary of the event. Live broadcast can be made more exciting and sophisticated (similar to a live TV studio broadcast), and requires algorithms for automatically making decisions about shot compositions and cuts (e.g., see [8, 13]). Automatically producing shorter movie summaries of virtual world events requires automatically recognizing main events, and also shot editing (e.g., both discussed in [11]). A challenge that came up is the need to combine shots taken from static or moving (either virtual or physical) cameras; however, such challenges are also addressed in traditional video editing.

We suggest that such hybrid events, augmented by famous virtual personas, can become increasingly popular. Above we discussed the ethical considerations for performing such an event, describing a general measure based on [19] and our measures taking them into account in the production of this event; such legaland ethical considerations are likely to become major issues of discussion.

Here we reported about a panel discussion in front of an audience, as part of an academic conference, but we can envision many other applications, including cultural events, entertainment, and education. Today, such events would typically use video conferencing software such as Zoom. However, multi-user VR may offer advantages over video conference, for both the participants as well as for the live audience; this is especially true for applications such as VR eSport events, music performances, or theater plays. Our hope is that others producing such events, combining multi-user VR in front of an audience, can learn from our experience. Moreover, we show how such events can be seamlessly enhanced by LLM-based virtual humans.

7 ACKNOWLEDGEMENTS

This work was partially supported by projects GuestXR (#101017884) and Socrates EU projects (#951930), which have received funding from the European Union's Horizon 2020 research and innovation program. MS is supported by the European Research Council Advanced Grant MoTIVE (#742989). The Einstein model was designed by Maya Shekel as part of Ronit Elyoseph's Ph.D. thesis research.

The development and production of this joint work within the 'Advanced Reality Lab' and 'EventLab'. We would like to credit to the stuff members that helped in the production: Virtual cinematographer: Maya Shekel; Event production: Alon Weizman (VRGo), Gal Yaar; physical moderator: Dr Jeremy Fogel; Panelists: Prof David Chalmers, Prof Mel Slater, Prof Doron Friedman, and Prof Albert Einstein.

REFERENCES

[1] Jean Baudrillard. 1994. *Simulacra and simulation.* University of Michigan press.
[2] Walter Benjamin. 1935. The Work of Art in the Age of Mechanical Reproduction, 1936.
[3] Michael Birnhack and Tal Morse. 2022. Digital remains: property or privacy? *International Journal of Law and Information Technology* 30, 3 (2022), 280–301.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[5] Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's GPT-2–how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774* (2019).

[6] Justine Cassell, Obed E Torres, and Scott Prevost. 1999. Turn taking versus discourse structure. *Machine conversations* (1999), 143–153.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Doron Friedman and Yishai A Feldman. 2006. Automated cinematic reasoning about camera behavior. *Expert Systems with Applications* 30, 4 (2006), 694–704.

[9] Doron Friedman and Béatrice S Hasler. 2016. The BEAMING proxy: towards virtual clones for communication. *Human Computer Confluence Transforming Human Experience Through Symbiotic Technologies* (2016), 156–174.

[10] Doron Friedman, Oren Salomon, and Béatrice S Hasler. 2013. Virtual substitute teacher: introducing the concept of a classroom proxy. *London, 28–29 November 2013 King's College London, UK* 186 (2013).

[11] Doron Friedman, A Shamir, YA Feldman, and Tsvi Dagan. 2004. Automated creation of movie summaries in interactive virtual environments. In *IEEE Virtual Reality 2004*. IEEE, 191–290.

[12] Doron Friedman and Peleg Tuchman. 2011. Virtual clones: Data-driven social navigation. In *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11*. Springer, 28–34.

[13] Rachel Heck, Michael Wallick, and Michael Gleicher. 2007. Virtual videography. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 3, 1 (2007), 4–es.

[14] Christoph Käding, Erik Rodner, Alexander Freytag, and Joachim Denzler. 2017. Fine-tuning deep neural networks in continuous learning scenarios. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13*. Springer, 588–605.

[15] Sameer Kishore, Xavi Navarro Muncunill, Pierre Bourdin, Keren Or-Berkers, Doron Friedman, and Mel Slater. 2016. Multi-destination beaming: apparently being in three places at once through robotic and virtual embodiment. *Frontiers in Robotics and AI* 3 (2016), 65.

[16] Donald Marinelli and Scott Stevens. 1998. Synthetic interviews: The art of creating a "Dyad" between humans and machine-based characters. In *Proceedings of the sixth ACM international conference on Multimedia: Technologies for interactive movies*. 11–16.

[17] Ramon Oliva, Alejandro Beacco, Jaime Gallego, Raul Gallego, and Mel Slater. 2023. The making of a Newspaper Interview in Virtual Reality: Realistic Avatars, Philosophy and Sushi. *IEEE CG&A* in press (2023).

[18] Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[19] Mel Slater, Cristina Gonzalez-Liencres, Patrick Haggard, Charlotte Vinkers, Rebecca Gregory-Clarke, Steve Jelley, Zillah Watson, Graham Breen, Raz Schwarz, William Steptoe, et al. 2020. The ethics of realism in virtual and augmented reality. *Frontiers in Virtual Reality* 1 (2020), 1.

[20] Anthony Steed, William Steptoe, Wole Oyekoya, Fabrizio Pece, Tim Weyrich, Jan Kautz, Doron Friedman, Angelika Peer, Massimiliano Solazzi, Franco Tecchia, et al. 2012. Beaming: an asymmetric telepresence system. *IEEE computer graphics and applications* 32, 6 (2012), 10–17.

[21] David Traum, Andrew Jones, Kia Hays, Heather Maio, Oleg Alexander, Ron Artstein, Paul Debevec, Alesia Gainer, Kallirroi Georgila, Kathleen Haase, et al. 2015. New Dimensions in Testimony: Digitally preserving a Holocaust survivor's interactive storytelling. In *Interactive Storytelling: 8th International Conference on Interactive Digital Storytelling, ICIDS 2015, Copenhagen, Denmark, November 30-December 4, 2015, Proceedings 8*. Springer, 269–281.

[22] Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology innovation management review* 9, 11 (2019).