nature machine intelligence

PERSPECTIVE

https://doi.org/10.1038/s42256-021-00417-9

AI-generated characters for supporting personalized learning and well-being

Pat Pataranutaporn[1], Valdemar Danry[1], Joanne Leong[1], Parinya Punpongsanon[2], Dan Novy[1], Pattie Maes[1] and Misha Sra[3] ✉

Advancements in machine learning have recently enabled the hyper-realistic synthesis of prose, images, audio and video data, in what is referred to as artificial intelligence (AI)-generated media. These techniques offer novel opportunities for creating interactions with digital portrayals of individuals that can inspire and intrigue us. AI-generated portrayals of characters can feature synthesized faces, bodies and voices of anyone, from a fictional character to a historical figure, or even a deceased family member. Although negative use cases of this technology have dominated the conversation so far, in this Perspective we highlight emerging positive use cases of AI-generated characters, specifically in supporting learning and well-being. We demonstrate an easy-to-use AI character generation pipeline to enable such outcomes and discuss ethical implications as well as the need for including traceability to help maintain trust in the generated media. As we look towards the future, we foresee generative media as a crucial part of the ever growing landscape of human–AI interaction.

The idea of computers generating content has been around since the 1950s. Some of the earliest attempts were focused on replicating human creativity by having computers generate visual art and music[1]. Unlike today's synthesized media, computer-generated content from the early era was far from realistic and easily distinguishable from that created by humans. It has taken decades and major leaps in artificial intelligence (AI) for generated content to reach a high level of realism.

Generative and discriminative models are two different approaches to machines learning from data. Although discriminative models can identify a person in an image, generative models can produce a new image of a person that has never existed before. Recent leaps in generative models include generative adversarial networks (GANs)[2]. Since their introduction, models for AI-generated media, such as GANs, have enabled the hyper-realistic synthesis of digital content, including the generation of photorealistic images, cloning of voices, animation of faces and translation of images from one form to another[3–6]. The GAN architecture includes two neural networks, a generator and a discriminator. The generator is responsible for generating new content that resembles the input data, while the discriminator's job is to differentiate the generated or fake output from the real data. The two networks compete and try to outperform each other in a closed-feedback loop, resulting in a gradual increase of the realism of the generated output.

GAN architectures can generate images of things that have never existed before, such as human faces[3,4]. However, StyleGAN is an example of a modifiable GAN that enables intuitive control of the facial details of generated images by separating high-level attributes like the identity of a person from low-level features such as hair or freckles, with few visible artefacts[4]. Researchers have also proposed an in-domain GAN inversion approach to enable the editing of GAN-generated images, allowing for de-aging or the addition of new facial expressions to existing photographs[7]. Meanwhile, transformers such as the ones used in the massive generative GPT-3 language model are already being shown to be successful for text-to-image generation[8].

In addition to the generation of new digital media, generative models like GANs can also manipulate existing media to enable applications such as video dubbing of foreign films[9], animating old images of historical figures[10] or, perhaps most notoriously, creating 'deepfake'[3] videos of people like world leaders, politicians and celebrities in which they are portrayed as saying or doing things they never did. However, the main goal of this Perspective is not to discuss the potential misuses of the technology, as these have been covered extensively in the literature (for example, refs. [3,11,12]). Instead, this Perspective aims to present beneficial applications of the technology, with a discussion of some of their societal implications and concerns (Social implications section).

The technology has found use for several beneficial applications. Currently, generative models are being leveraged across different industries including entertainment, customer services and marketing. Examples include mobile device apps (for example, Reface[13]) that enable users to humorously swap their faces in video clips and GIFs to share with friends. Other startups allow users to create AI-generated photorealistic virtual assistants or to perform face replacement in videos or live streams[14]. Virtual characters like Li'l Miquela[15] are already exceedingly popular, with millions of followers on Instagram. Beyond images and videos, a model generating realistic lip movements in videos based on voice recordings has recently been made public, allowing anyone to translate recorded lectures, announcements and movies into another language, without any noticeable discrepancy between speaker lip movement and voice overdubs[16].

Given the recent advancements in generative AI, we suggest that AI-generated characters have the potential to achieve promising results in learning and healthcare, but more research is needed to determine valuable opportunities and understand their limitations. In this Perspective we present emerging use cases of AI-generated characters and discuss their potential in education and well-being (Fig. 1). Additionally, we demonstrate and make available an easy-to-use AI character generation pipeline (https://github.com/mitmedialab/AI-generated-characters) based on publicly available

[1]Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Osaka University, Toyonaka, Japan. [3]University of California, Santa Barbara, CA, USA. ✉e-mail: sra@cs.ucsb.edu

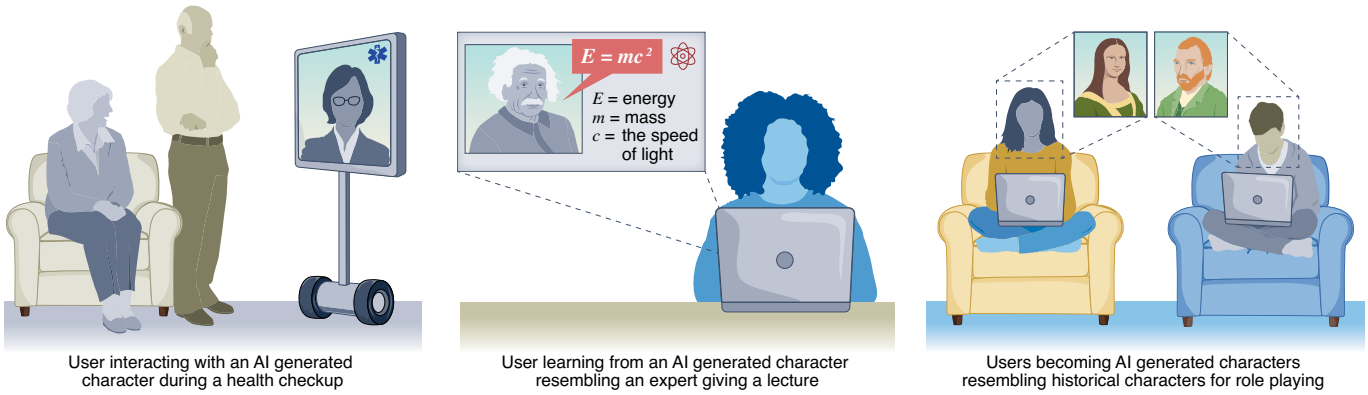

Fig. 1 | Example applications of AI-generated characters that resemble contemporary and historical figures. Users are shown interacting with an AI-generated character during a health check (left), learning from an AI-generated character resembling an expert in a lecture (middle) and becoming AI-generated characters that resemble historical characters for role playing (right).

code. We highlight some of the ethical implications of AI-generated characters, as well as the need for including traceability to help maintain trust in digital media. As we look towards the future, we foresee generative media as being part of an ever growing landscape of connections between humans and AI.

AI-generated characters

In this Perspective, we define AI-generated characters as digital representations of a person created by machine learning algorithms that are made to look, sound and behave realistically without necessarily being malicious (like deepfakes). To contextualize and position AI-generated characters in the broader research field of humanoid computational characters, we construct a design space with three axes: realism, interactivity and embodiment (Fig. 2). AI-generated characters are virtual, exhibit varying degrees of interactivity, and can be more realistic than hand-crafted characters (like three-dimensional (3D) or CGI/VFX (computer-generated imagery/visual effects) characters). The increased realism is due to the fact that AI-generated characters can leverage a large dataset of human images and videos to create output that resembles a photorealistic input. This stands in contrast to manually crafting a CGI character, which relies heavily on the creator's expertise in manipulating individual 3D parameters such as eye shape or size to satisfy their subjective evaluation of realism. AI-generated characters are—given their virtual embodiment—distinct from physically embodied characters such as humanoid robots (Fig. 2). In terms of proposed applications, we identify two sub-types of AI-generated character: (1) a user becoming an AI-generated character with a face swap and (2) a user interacting with an AI-generated character, which is controlled by either another human or an intelligent system. We see both of these sub-types as having unique strengths that enable novel applications in a variety of settings, as exemplified in the following.

The first sub-type allows a person to wear the face of a generated character that can help the user imagine themselves as someone else, and motivate them to consider counterfactual scenarios, gain new perspectives and allow them to experiment with ideas through role playing. Virtually being someone else has been shown to increase objective thinking capabilities[17], strengthen engagement[18] and demote discriminatory biases[19], and we see this sub-type leading to similar effects. The second sub-type allows users to interact with other users wearing an AI-generated face or with embodied intelligent systems. The latter can serve as motivators by depicting a personally meaningful mentor, adviser or supporter, exhibiting their facial features, voice and mannerisms. Easy availability

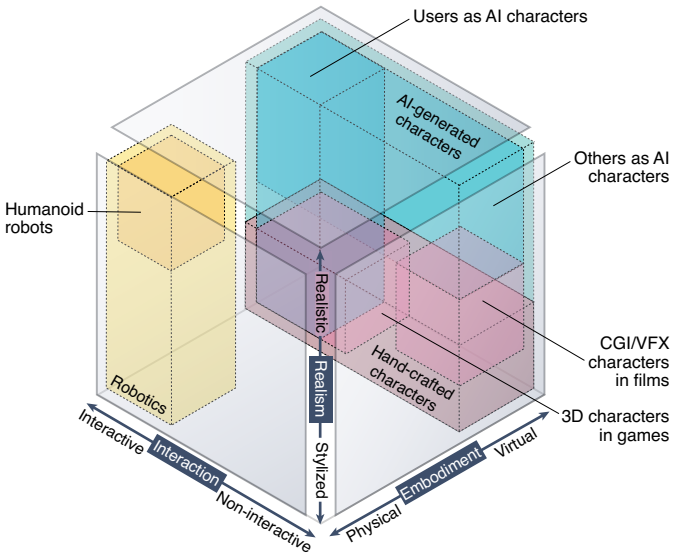

Fig. 2 | AI-generated characters as a domain can be characterized along three axes: realism, embodiment and interactivity. Within these axes, AI-generated characters can be compared with other domains, such as humanoid robots, CGI/VFX characters in films, and 3D-characters in games.

of these characters has the potential to broaden benefits that may come from unrestricted access to mentors of an individual's choosing[20]. AI-generated characters can also provide access to educational content, regardless of location or resources. For example, a virtual character for teaching sustainability to grade-school children was successfully tested in New Zealand in response to teacher shortages[21]. Health messaging, counselling, engagement with present-day topics and many other aspects of life can be supported by generative AI and can be personalized for each individual in ways not easily possible with real humans in those roles.

Both sub-types of AI character can be made accessible in multiple forms (video, 3D, audio) and through multiple devices, such as smartphones, laptops or augmented and virtual-reality headsets, which can provide new opportunities for engagement and interaction with the generated representations. The character portrayals can feature different degrees of fidelity and interactivity, from full bodies to voice- or text-only manifestations.

AI-generated characters for education

Students spend ~15,000 h learning in the first two decades of their lives. Thus, learning environments have an enormous impact on the individual and, in turn, on society[22]. AI characters can be used to create compelling learning material for all ages, from delivery of content in the classroom to engagement with content at locations like museums, historical monuments or even in nature. By using state-of-the-art algorithms, prominent historical, modern-day or fictional figures can be brought to life to engage learners with 'lived' experiences of scientists making their discoveries, historical figures narrating battles or painters discussing their inspiration and process.

AI-generated characters as virtual instructors. An ideal learning environment is one that succeeds in promoting motivation for learning, including through the optimization of the student–teacher relationship[23]. Research has shown that access to technology does not automatically equate with improved motivation or learning outcomes. Instead, the way in which teachers integrate the technology into their lesson plans plays an important role in motivation, along with student psychological processes and contextual factors involved in learning[24]. In a learning context, various studies have shown that character appearance can positively impact learners' behaviours, attitudes, and motivation[25]. Learning experiences blended with fictional characters and narratives have also been shown to enhance attention and engagement[26]. This was demonstrated in early 2020 when a professor saw a spike in student enthusiasm after teaching his online classes as an anime character[27]. His students became particularly fond of the anime character and he noticed an overall increase in student engagement as well as noticeable improvements in retention for both his undergraduate and graduate courses.

Such immersive opportunities support a variety of novel scenarios for lifelong learning, whether they involve learning about art from famous artists or learning to cook from the best chefs in the world. For example, in the 'Dalí Lives' exhibition[10], a generated version of the famous artist talks to visitors and even takes selfies with them in the museum. This interactive installation reportedly helped people to more closely connect with Dalí and his artwork. It is easy to imagine other artists coming to 'life' through generated versions of themselves, showcasing the potential of AI-generated characters to motivate learning both in and outside the classroom.

Students embodying AI-generated characters. The usage of AI-generated characters for learning can be further expanded, with not only teachers but also students playing different roles. For example, a student can become Einstein and try to outline his theory of relativity or, during a history lesson, multiple students can embody the founding fathers of the United States and participate in the writing of the constitution. Researchers have demonstrated that embodying a character in virtual reality can positively influence people's behaviours and abilities in multiple ways, from supporting expressivity[28] to mitigating negative stereotypes[29], and enhancing problem-solving skills. Furthermore, researchers have studied engineering students who embodied 'inventor' characters during collaborative brainstorming tasks in virtual reality, and found them to have greater fluency and originality of ideas compared to those under control conditions[30]. Researchers also found that virtually embodying an Albert Einstein character in virtual reality improved cognitive task performance and decreased age bias[31]. Motivated by research in virtual reality, researchers have begun to study the effects that real-time camera filters may have on embodiment and cognitive functions such as creativity[32].

AI-generated characters as peers. Finally, beyond having AI-generated characters as instructors or students role playing for a more immersive learning experience, we also envision the use of generated characters as peers in learning support groups. Social rewards, such as praise from others, have been shown to promote greater motivation in children[33], improve academic performance[34] and increase self-efficacy[35]. Praise from artificial entities such as robots and virtual agents has similarly been shown to enhance human motivation and task performance[36]. Researchers have also shown that a robot designed as a social agent that interacts with children as a peer (not as a tutor or teacher) can help students organically improve their language skills through social activities such as storytelling games, during which the agent introduces new vocabulary words and models good story narration skills[37]. These examples suggest that AI-generated characters may be able to enhance motivation, not only by their appearance but also through their interaction and feedback.

AI-generated characters to support health and well-being

Personal well-being is defined in the literature as a happy, satisfactory and desirable state of being that includes physical, psychological, emotional, mental, spiritual, social and subjective well-being[38]. We believe that AI-generated characters, in the future, will support several aspects of well-being by augmenting the assistance provided by coaches, therapists or counsellors, especially in resource-constrained environments.

Digital health counselling. Only ~45% of individuals with mental health issues in the United States received treatment such as counselling in 2019[39]. With warnings of a critical shortage of affordable therapists and psychiatrists, unsurprisingly, digital mental health has become a multi-million-dollar industry, with more than 10,000 apps[40]. Woebot, a conversational agent (or chatbot) introduced in 2017, is one of the few AI-based mental health apps to deploy the principles of cognitive behavioural therapy (CBT), which is commonly used to treat anxiety and depression. With four in ten US adults reporting anxiety or depression in 2020, Woebot has seen its number of daily users double to tens of thousands during the COVID-19 pandemic (March 2020–June 2021)[40].

Conversations with agents such as Woebot have been shown to resemble human conversations[41] and result in emotional bonds with interactees. Examples of conversational agents include (but are not limited to) text-based chatbots, which can engage in casual conversations; voice-based conversational interfaces like Apple Siri, Google Assistant or Amazon Alexa, which respond to questions and instructions; and embodied conversational agents, which involve a virtual character (for example, avatar, virtual agent or photorealistic character) simulating face-to-face conversation with verbal and nonverbal behaviours[42]. The potential of text-based conversational agents for supporting mental health and well-being has already been shown to be significant in previous work: randomized control trials of the use of conversational agents (chatbots and/or embodied agents) in psychiatry and mental health counselling have demonstrated significant reductions in psychological distress compared with inactive control groups[43]. As another form of social support, the potential benefits of conversational agents in mitigating loneliness—a growing public health issue—have been suggested by researchers[44]. We can envision embodied chatbots or conversational AI agents that may be able to provide support through voice-based conversations and facial expressions as a further evolution of the text-messaging-based conversations of today's chatbots. By combining conversational agents with AI-generated characters, there is potential to enable deeper personalization and increased trust. Earlier research has shown that characteristics of doctors and therapists, such as their gender, can lead to significant differences in treatment outcomes. Indeed, one study found that patient satisfaction with the therapeutic relationship was higher when people were paired with same-gender therapists[45], while another study found personal characteristics to be strongly associated with trust

in physicians[46]. Similarly, interacting with computer-generated characters has also been found to establish better levels of trust and support open communication. In a study on people with eating disorders, a cartoon-like avatar of the therapist was shown to facilitate greater openness from participants[47].

Empirical evidence has shown the potential of conversational agents to induce positive behavioural and cognitive changes[48]. By combining conversational agents with AI-generated characters, personalized fitness apps could enable users to interact with and be guided by a personalized AI-generated coach or exercise buddy that resembles someone they find motivating or look up to as a role model[20].

Exposure therapy. Virtual reality has shown to be effective in inducing emotional responses that match real-world experiences, making it extremely valuable in exposure treatment. In virtual environments, patients experience similar physiological symptoms and fear as they might in equivalent real-life situations, thereby facilitating the habituation process[49]. For example, one virtual reality-based study found that users exposed to age-progressed renderings of themselves had an 'increased tendency to prefer delayed monetary rewards over immediate ones'[50]. Similarly, patients exposed to familiar experiences such as prior relationships were shown to maximize functional abilities in a study of elders with dementia[51].

Based on these results, we believe that AI-generated characters could be useful in exposure therapy. Exposing someone to AI-generated versions of themselves, for example, younger versions of themselves, can have a positive impact on mental health and has already been shown to improve alertness, active participation and a general sense of well-being[52].

Living memories. Every human culture has developed practices and rituals associated with responding to death and signifying its importance for the living[53]. The digital world is increasingly intersecting with these practices. For example, digital platforms like Facebook already enable people to revisit deceased friends' and families' profiles as 'digital graves'. In the last year or so, Zoom has also been streaming memorial services and funerals for family members unable to attend in person due to COVID-19[54]. Research in human–computer interaction has traditionally explored the full range of the human lifespan, with a new call in the past decade for technologies to support practices such as 'collaborative acts of remembrance, bequeathing of digital data, or group reflection on the digital residual of a life'[53,55]. We believe that the concept of living memories through AI-generated characters could provide a new form of 'technological heirloom' and a new way for reflection and meaning-making[53], adding to existing objects of memorabilia of the deceased, such as photographs, videos and other objects.

A recent example includes an AI developer creating a chatbot of her deceased friend[56]. After sharing the chatbot with other friends of the deceased, she claimed that, for many people, interacting with the bot had a therapeutic effect and that it deeply impacted the lives of people who knew him. His mother told her, 'I am getting to know him more. This gives the illusion that he is here now.' Another example is The Deep Nostalgia service, offered by a genealogy company MyHeritage, which deploys an AI-driven video reenactment technology to animate the faces of departed family members and ancestors in still photos and creates high-quality, realistic moving videos (https://www.myheritage.com/deep-nostalgia). The service Eterni.me has a similar goal of creating and preserving memories through interactive conversations with the departed, and has garnered over 30,000 people waiting to be digitally immortalized for friends and family—including patients with terminal diseases such as cancer and Alzheimers[57]. Photogrammetry was used in another recent recreation of a deceased child[58] to allow the parents to remember their son and cope with their traumatic loss. Technology now enables the preservation of a large variety of data, from text and photographs to videos and voice data. Although it might seem controversial to 'digitally preserve' individuals beyond death as was explored in fictional works such as *Black Mirror*[59], 'these AI systems may support the grieving process more effectively and speed up the stages of the process which include denial, anger, bargaining, depression and acceptance'[60].

Assistive technologies. Generative media can enable people with a facial deformity or speech inability resulting from injury or a disorder such as amyotrophic lateral sclerosis (ALS) to restore their verbal or facial expressiveness in digitally mediated interactions. Researchers have recently developed non-invasive, real-time silent speech systems to assist patients with speech disorders such as multiple sclerosis and ALS to communicate in natural language merely by articulating words or sentences in the mouth without producing any sounds[61]. Although this work enables patients to regain close to natural sounding speech through a standard voice synthesizer, it does so by using a non-personal generic voice. AI-generated face and voice replication could enable them to not only regain speech but also their own voice. This idea was realized by an engineer diagnosed with motor neurone disease. In anticipation of muscle deterioration caused by the disease, he developed a life-like avatar with his facial features and voice, trained on data recorded prior to the progression of his illness[62], that he now uses for digital communication.

AI-generated characters from text, audio and video

We present a unified pipeline for enabling educators and individuals to easily generate talking characters that can be useful for educational, health and other applications. The pipeline allows for the easy creation of realistic content based on real and fictional humanoid characters with facial gestures, voice and motion. It combines state-of-the-art generative AI models that convert text to audio, audio to video, and video to video, and can be used to create a variety of audio and video outputs based on the available data.

The following models are used: VOCA[63] (Voice Operated Character Animation), FLAME[64] (Faces Learned with an Articulated Model and Expressions), Speech-Driven Facial Animation[65] and First Order Motion Model[66]. After the generation process, our unified pipeline marks the generated results with a traceable watermark to help people identify that the data are generated, learn about how the output was generated, and distinguish it from authentic video content. Our pipeline differs from existing consumer apps in allowing any type of input medium for generating videos and real-time facial filters while remaining open-source. This lends educators, therapists, health officials and others a higher degree of autonomy in the usage and creation of AI-generated characters.

Video-to-Video. To generate the video output, the user provides a template video and an image of the target character's face. The algorithm maps features from the video template to the target character to create a realistic video. To map the features, our pipeline uses the state-of-the-art First Order Motion Model generator network[66] (https://github.com/alievk/avatarify-python), which utilizes a self-supervised learning algorithm to determine the local transformation of facial animation to generate realistic complex motion in the output (https://github.com/alievk/avatarify-python). To generate a facial filter for real-time use, the user's webcam stream is used as input (Fig. 4 and top of Fig. 3).

Voice-to-Video. Voice is a critical element of generated output given that watching a generated video without audio is less likely to be perceived as realistic. To add audio to the generated video, the user provides a recorded voice clip to drive the generated character's speech. As shown in Fig. 3, we explore two pathways for converting the voice clip input into intermediary facial animation before generating the final video and audio output.

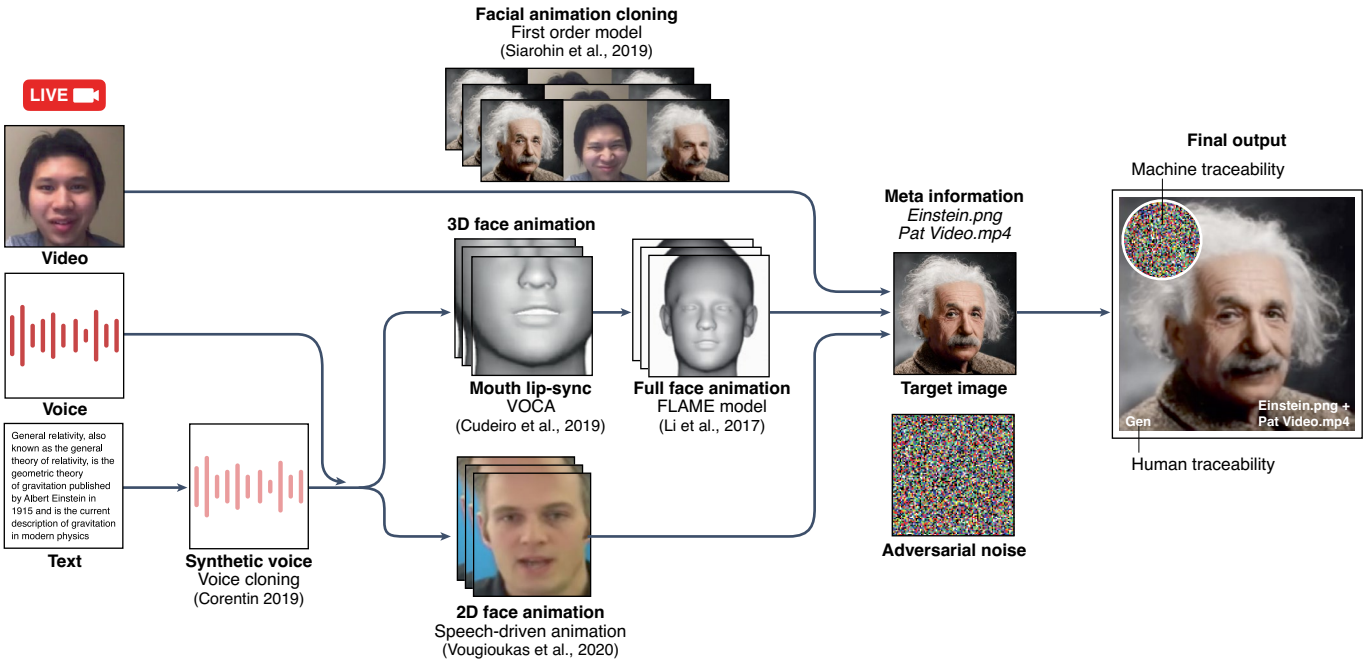

Fig. 3 | Our unified pipeline allows users to provide video, voice or text as inputs to generate videos and real-time facial filters. The output is watermarked for traceability (Traceability section) for detection both by humans and algorithms. Image of face at bottom centre reproduced with permission from ref. [65], Springer Nature. Image of Albert Einstein taken from Pixabay. The following references are cited in the figure: refs. [63–66,100].

Of the two approaches, the first approach creates a 3D intermediary talking head that can automatically lip-sync with the given voice clip. To do this we first use VOCA, a deep neural network trained on 4D scans of human faces and synchronized speech from speakers, to automatically generate lip-sync animation on a neutral 3D face[63]. Next, we use FLAME, a generalized 3D facial model, to animate the neutral face generated in the previous step[64]. The FLAME model allows the creator to manually customize the expressions and movements of the generated 3D face, while VOCA automatically synchronizes the mouth with the given voice clip.

The second approach utilizes an end-to-end speech-driven facial animation system based on a temporal GAN[65]. The temporal GAN relies on a generator and three discriminators (frame, sequence and synchronization discriminators) that drive the generation of an auto-lip-sync talking head using only a still 2D image of a person and a voice clip as input. The advantage of the second approach is its capability in synthesizing natural facial expressions such as blinks and eyebrow movements without manual customization.

Text-to-Video. Instead of using a recorded voice as input, text can be used for the generation of a voice clip and further use of our pipeline. If the user does not want to use their own voice or does not have the character's pre-recorded audio, they can simply enter a sentence, phrase or paragraph to drive the generated character's speech. For example, a generated Einstein could speak in Einstein's own voice or Elsa's voice or the teacher's voice who is puppeteering Einstein's face. To synthesize the generated character's speech, we use a state-of-the-art publicly available real-time voice cloning algorithm[5,67] that enables using a few seconds of reference audio to synthesize new speech in that speaker's voice, using the text data. After synthesis, the audio is then used for the intermediary facial animation synthesis step described in the Voice-to-Video section.

Traceability. Given the potential of generative AI to mislead the consumer of the content, it is important to leave a trace to enable people to identify, investigate and debunk the authenticity of the generated media[68]. To ensure the safe and ethical use of our pipeline, we include human and machine traceability methods. Our human traceability technique is inspired by fabrication detection techniques in other media paradigms (like text and video) and involves incorporating perceivable traces like signatures of authorship, distinguishable appearance or small editing artefacts into the generated media[68]. Machine traceability, on the other hand, involves incorporating traces imperceptible to humans, like invisible noise that only machines can detect. In our pipeline, we implemented human traceability by visually marking the generated content (in our example case with 'Gen' for 'generated' and the names of the target image and driving video files, 'Einstein.png' and 'Pat_Video.mp4') and machine traceability by using adversarial noise[69].

Our implementation of traceability contrasts with typically used methods such as deep-learning-based detectors that act as post factum detectors after the content has already been created[3]. Because researchers worry that detection algorithms will always remain a step behind[3] given the rapid pace of technological advancements, our approach differs from typical detection methods by ensuring traceability through marking the content during the generation process.

Pipeline in action. To demonstrate the use of our pipeline, we experimented with generative characters in online classes at Massachusetts Institute of Technology (MIT). First, we collaborated with an MIT professor for his musical interfaces class by creating a real-time AI-synthesized version (sub-type 1) of Johann Sebastian Bach (Fig. 4b, bottom row), who had a live conversation with the renowned cellist and guest speaker, Yo-Yo Ma. Seeing their musical hero brought to life enthralled all class attendees, and this was followed by a lively discussion of Bach's work. We then created a video of Einstein from his audio recording of the essay titled *The Common Language of Science* (Fig. 3) and showed the video in another class at MIT to demonstrate how easy it is to convert information that is available only in audio or written format into a video format that engages more senses and thus may be more engaging for learners.

Fig. 4 | Using AI-generated characters in a virtual classroom setting. a, One of the authors of this Perspective (bottom row) becoming an AI-generated George Washington (top row). **b**, One of the authors of this Perspective embodying Johann Sebastian Bach (bottom row) interacting with Yo-Yo Ma (top row) in the online class 'ARTS@ML' in autumn 2020. Panel **a** (top), George Washington portrait by Gilbert Stuart, image courtesy of the Clark Art Institute (clarkart.edu). Panel **b** (top), image used with permission of Yo-Yo Ma; (bottom) image of Johann Sebastian Bach taken from Wikipedia.

Social implications

Historically, most new technologies have raised ethical concerns of various sorts, from the quaint (beliefs about women's bodies not being able to go at 50 m.p.h. in trains[70]) to the fearful (beliefs about the first telephones destroying face-to-face communication[71]) and the true (cars will change the landscape)[72]. Some ethical concerns have resulted in global policy efforts for the control of new technologies (such as nuclear power), while others have resulted in smaller local regulations and laws.

With the rapid growth of AI technologies, concerns around privacy and surveillance, manipulation of behaviour, innate biases, as well as responsibility and the delegation of decision making, have all been raised[72].

In this Perspective we envision how recent progress in AI-generated characters can make it easy to create and animate realistic digital portrayals of people for novel applications. Although this technology enables exciting possibilities in areas such as learning, health and well-being, AI-generated media can and have been used for malicious purposes. For example, celebrity faces have been swapped into pornographic videos[11], the US president has been made to support criminal deception[73], and the general population has been targeted with realistic fake news and harassment (or 'trolling') at levels beyond the capability of human detection[74]. Given these real challenges of privacy breaches, defamation, manipulation and misinformation, acknowledging and addressing ethical and legal concerns are essential for positively integrating this technology in society.

Ethical considerations. Discussion regarding the ethics of AI-generated media is an ongoing, expansive conversation happening across different scales, from personal usage to national policies[12,75–77]. The goal of this discussion is to present potential ways in which the technology can be misused and to promote considerations on which to reflect when designing AI-generated characters. In this section, we acknowledge the complexity of the potential for misuse and highlight some concerns as cautionary items that we encourage people to consider when using AI-generated characters in health and education contexts. We further acknowledge that the concerns presented here might not be an exhaustive list of all potential misuses of AI-generated characters.

Misportrayal of a character or event. Our ability to acquire knowledge depends on information from sources that we consider reliable[77]. Although generative AI allows us to synthesize media that assist and engage people in acquiring new knowledge, such media can also be misused by inaccurately depicting figures and distorting expert opinions. This is especially important in high-impact domains such as education and health, where misportrayal can adversely affect individuals. AI-generated characters could be made that portray a historical and trusted figure like Albert Einstein stating that the Earth is flat, or to alter and distort historical events, as demonstrated with the 2020 installation 'In Event of Moon Disaster'[78]. This installation, with the goal of creating public awareness around AI-generated media and misinformation, falsely depicts Nixon giving a speech stating that the Apollo 11 crew passed away on their moon landing mission. Generated media can be used to depict encounters, conversations, behaviours and events that never happened. Such misportrayals of historical events or trusted figures have the risk of spreading misinformation and destabilizing trust in media[77].

Promoting harmful behaviour. Beyond the misportrayal of characters, another potential misuse of AI-generated characters could be having them support and endorse harmful behaviour (https://www.pnas.org/content/118/47/e2114388118). For example, prominent official health figures could be made to speak against COVID-19 vaccinations. In one case study, researchers followed a Twitter account imitating a Disney princess endorsing self-harm to exemplify the negative effects of fake representations of beloved characters[79]. Among other things, the account shared and praised pictures of the bodies of people with eating disorders and self-injuries like scars from cutting themselves. With much content on video platforms like YouTube being consumed by children[80], recognizable characters promoting harmful behaviour could have severe consequences, such as the development of eating disorders and suicidal ideation. Furthermore, AI-generated characters can unintentionally cause harm. Imagine, for instance, a classroom using synthetic characters to allow students to debate each other as presidential candidates for an upcoming election. If these student depictions somehow find their way into the public domain, they might be perceived as true or real by some individuals.

Causing dependence from overuse. Interactions with AI characters can be personalized to increase engagement and motivation in learning and information-consumption contexts. In the future, characters could be hyper-personalized to such a degree that they are preferred over real human instruction and interactions. Social media use is already increasingly correlated with perceived social isolation[81]. Reduced human social relationships adversely affect mental health, physical health and mortality risk[82]. It is possible that excessive interaction with digital humans could ultimately negatively impact learning and communication in ways similar to how excessive video-game playing and computer use are shown to be associated with depression[83]. One solution would be to limit interactions with AI characters to situations that demand it, instead of replacing interactions that could be performed with real humans (for example, filling the teacher gap in developing nations or providing support to users between sessions with a human therapist).

Supplanting human relationships. Although we have presented applications of AI-generated characters for teaching, coaching and therapy, we would like to stress that they should be used for augmenting current practices rather than replacing them. For example, although it might be economically motivating to replace teachers with AI systems, research has shown student–teacher relationships to be a key factor in creating an environment for positive student development and learning[23,34]. Substituting teachers with AI-generated characters might thus jeopardize the ability of the students to form genuine relationships with their teachers and mentors, which in turn may impact learning motivation and cause disruptive behaviour[84]. Moreover, current video conferencing systems for online learning have been shown to lack emotional attachment and lead to ineffective learning experiences[85]. Although human resources like teachers and therapists are not always available in low-resource communities, we do, based on these findings, consider it important that AI-generated characters in their current state be used mainly as a supplementary tool rather than a replacement.

Post-mortem AI-generated characters. Creating an AI-generated character based on a deceased person is a highly controversial use case, with complex and delicate personal and moral implications. On the one hand, the idea of externalizing and imprinting a person's characteristics, knowledge, stories and wisdom into an artefact (for example, personal diaries, old recipe books, piece of clothing or accessories) that can be passed on from generation to generation is a defining feature of humans as a species. We create artefacts to remember or be remembered by people that are important to us. On the other hand, the idea of using technology to convert a person into computer code and patterns of data can be viewed as reductive and dehumanizing. This moral dilemma perhaps depends on the interpretation, utility and cultural context of the person analysing the issue. A review of designs dealing with death highlighted that digital approaches such as AI-generated characters resembling a deceased person can provide an opportunity to deal with grief that can empower a user and help them cope with their loss[55]. Today, Eterni.me and Replika already build chatbots that enable people to be given 'life' after death[57]. Eterni.me[57] specializes in post-mortem AI-generated characters and has garnered tremendous interest from cancer patients wishing to have an AI copy of themselves left behind. Future research needs to investigate the implications and consequences of such an interactive modality for memory preservation. Questions remain about whether the generated characters can stay authentic to the deceased and whether they may be confused with the real person. In the future, we may have wills that not only deal with an individual's physical possessions but also their digital identity, with declarations on whether one wants his or her data to be used to 'live on' as an AI-generated character and in what manner. However, questions regarding digital replicas of deceased individuals are a matter of discussion for both the surviving family and friends, as well as the deceased individual.

Legal considerations. In this section we discuss the legal considerations of AI-generated characters, which involve the rights and consequences for a number of individuals: the target individual whose identity is used (and who may be alive or deceased), the individual(s) whose data are used in training the AI models, the individual(s) building and training the model for AI character generation, the people using the model to create characters and the people who will view or interact with the AI-generated characters made by someone else.

Violation of an individual's rights of privacy and publicity. Actor Kirsten Bell, a victim of deepfake pornography, has said 'We're having this gigantic conversation about consent, and I don't consent, so that's why it's not okay… even if it's labelled as, "This is not actually her", it's hard to think about that'[86]. Using AI-generated media to put someone's face in a pornography video or images is a sexual-privacy invasion, which can have profound effects. Victims report experiencing deep psychological distress, including anxiety, depression, loss of appetite and suicidal ideation[87]. In a recent study that assesses public attitudes towards deepfakes, subjects viewed non-consensually created pornographic deepfake videos as extremely harmful and overwhelmingly wanted to criminalize them. Labeling pornographic deepfakes as fictional did not mitigate the videos' perceived wrongfulness. By contrast, non-pornographic deepfakes were viewed as substantially less wrongful when they were labeled as fictional or did not depict inherently defamatory conduct like illegal drug use[88]. Furthermore, labelling the generated pornographic videos as fiction does not meaningfully remove this harm, as the target's identity is still being appropriated[88]. We clearly see that using AI-generated media without consent is unethical for negative applications. However, there is still ongoing debate about the necessity of consent when it comes to positive use cases, such as the use of David Beckham for a malaria information campaign[89] or the use of a historical character such as Dalí to interact with museum-goers[10]. Does consent change if an individual passes away? Does their legal identity evaporate or does it lie with their estate? Who provides consent for an historical figure? As these considerations may vary across context and applications, they will most likely continue to evolve as AI-generated characters enter mainstream modes of communication.

Liability. If an AI-generated character misportrays or causes harm, who is to blame? Is it the person interacting with the AI-generated character? Is it the programmer who build the AI? Or is it the person who is portrayed? Determining the liability of AI-generated characters is an important topic, as teachers, health workers, agencies and others need to spread awareness of potential liability for claims related to AI-generated content. If an AI-generated character in some way violates someone's privacy, defames them or spreads misinformation, it is essential to determine who should be held responsible. Although some current laws apply, other laws and regulations are unclear and inadequate to deal with such situations[90].

In a recently published article on the ownership of AI-generated content, it is argued that, to reach guiding principles of liability, we must distinguish the contributions of different individuals in the creation process with regards to 'creativity, labour or intellectual effort'[90]. For example, in cases where control over the AI-generated content is strictly bounded by the programmer who created the algorithm, the programmer is—more often than the user—considered liable for the content (with variations existing between jurisdictions)[90]. Conversely, in cases where a great amount of freedom in content generation is given, users might be considered more liable than programmers.

AI-generated characters add an extra layer to the challenge of liability for generated media, as they are often very realistic (in contrast

to artistic), unlabelled and illusory by nature, making it hard to even determine if the character has been programmatically created at all. In this Perspective we have argued that the labelling of AI-generated media could enable some traceability regarding who is involved in the generation process, to what extent and what sources were used to generate the output. Other researchers have made similar arguments, stating that such labelling can greatly help mitigate liability issues by clearly marking something as AI-generated[91].

But can liability be resolved that easily? Numerous psychological studies of people's susceptibility to misinformation have, for example, shown that continuous exposure to clearly labelled misinformation can still have adverse effects on an individual's beliefs[92]. Similarly, using AI-generated media in learning and health, one should be encouraged to go beyond traceability and also think about the social and individual impact a particular piece of generated content might have.

Conclusion

In this Perspective, we have discussed beneficial applications of AI-generated characters in education and well-being, an easy-to-use pipeline for integrating them into these fields, as well as a discussion on the societal implications of using AI-generated characters. We foresee a future where generative AI will become ubiquitous in our daily lives, as part of the ever growing human–AI interaction landscape, but this is not without challenges[93]. Beyond the education and well-being domains, these technologies are likely to be used for a positive impact in many other areas, including entertainment, creativity and security. Although current algorithms are far from perfect, they open up new creative opportunities for artists and filmmakers without requiring Hollywood-sized budgets and timelines. Although CGI is commonly used in Hollywood to bring deceased actors back to life or to de-age stars, fans have been experimenting with replacing the CGI characters in movies (for example, Star Wars) with convincing AI-generated versions. With these fan versions gaining popularity, and Disney Studios themselves exploring neural face swapping[94] technologies, big movie productions may shift from manual and time-intensive CGI to algorithmic film making. As AI-generated characters become more familiar to the public, we also anticipate them to integrate with innovations in robotics[95], synthetic biology[96] and manufacturing[97] to support the presence of generative AI in the physical world. Companies are already working towards that vision by making ubiquitous characters that are deeply integrated with our lives[98,99]. However, with these opportunities come potential misuses. Challenges around AI-generated synthetic media in general involve multidimensional issues that touch on the notions of privacy, freedom of speech and the fundamental idea of human identity. AI-generated characters present the potential for misportraying people and events, motivating harmful behaviour, causing dependence from overuse, replacing social connections and promoting the digital existence of the deceased. Integrating this technology into social practices requires careful and critical consideration of issues such as privacy, liability and traceability. This will require cooperation between governments, industries, researchers and academic institutions to develop new regulatory mechanisms and safeguards that can address the technical, ethical and legal challenges. Our inevitable future with AI-generated characters will require us to rethink the fundamentals of human identity, its formation, its safeguarding and its role in society.

Received: 30 July 2020; Accepted: 26 October 2021;
Published online: 15 December 2021

References

1. Boden, M. A. & Edmonds, E. A. What is generative art? *Digital Creativity* **20**, 21–46 (2009).
2. Goodfellow, I. et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems* 2672–2680 (NIPS, 2014).
3. Mirsky, Y. & Lee, W. The creation and detection of deepfakes: a survey. *ACM Comput. Surveys* **54**, 1–41 (2021).
4. Karras, T. et al. Analyzing and improving the image quality of StyleGAN. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 8110–8119 (IEEE, 2020).
5. Zhang, Y. et al. Learning to speak fluently in a foreign language: multilingual speech synthesis and cross-language voice cloning. Preprint at https://arxiv.org/abs/1907.04448 (2019).
6. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 1125–1134 (IEEE, 2017).
7. Zhu, J., Shen, Y., Zhao, D. & Zhou, B. In-domain GAN inversion for real image editing. Preprint at https://arxiv.org/abs/2004.00049 (2020).
8. Ramesh, A. Zero-shot text-to-image generation. In *Proc. 38th International Conference on Machine Learning* Vol. 139, 8821–8831 (PMLR, 2021).
9. Takahashi, D. Deepdub uses AI to localize voice dubbing for foreign language films. *Venture Beat* (16 December 2020); https://venturebeat.com/2020/12/16/deepdub-uses-ai-to-localize-dubbing-for-foreign-language-films/
10. Dalí lives (via artificial intelligence) (Salvador Dali Museum, 2020); https://thedali.org/exhibit/dali-lives/
11. Westerlund, M. The emergence of deepfake technology: a review. *Technol. Innov. Manag. Rev.* **9**, 40–53 (2019).
12. McCammon, M. N. in *The Handbook of Communication Rights, Law and Ethics* Ch. 24 (Wiley, 2021); https://doi.org/10.1002/9781119719564.ch24
13. ReFace. Swap. Share. Hype. https://reface.app/ (accessed 10 July 2020).
14. Pinscreen. The most advanced AI-driven virtual avatars. https://www.pinscreen.com/ (accessed 8 October 2020).
15. Emilia, P. Who is Lil Miquela, the digital avatar instagram influencer? https://www.thecut.com/2018/05/lil-miquela-digital-avatar-instagram-influencer.html (accessed 24 December 2020).
16. Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P. & Jawahar, C. A lip sync expert is all you need for speech to lip generation in the wild. In *Proc. 28th ACM International Conference on Multimedia* 484-492 (ACM, 2020); https://doi.org/10.1145/3394171.3413532
17. Osimo, S. A., Pizarro, R., Spanlang, B. & Slater, M. Conversations between self and self as Sigmund Freud—a virtual body ownership paradigm for self counselling. *Sci. Rep.* **5**, 13899 (2015).
18. Slater, M. et al. Virtually being Lenin enhances presence and engagement in a scene from the Russian revolution. *Front. Robot. AI* **5**, 91 (2018).
19. Peck, T. C., Seinfeld, S., Aglioti, S. M. & Slater, M. Putting yourself in the skin of a black avatar reduces implicit racial bias. *Conscious. Cogn.* **22**, 779–787 (2013).
20. Pataranutaporn, P., Vega Gálvez, T., Yoo, L., Chhetri, A. & Maes, P. Wearable wisdom: an intelligent audio-based system for mediating wisdom and advice. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* 1–8 (ACM, 2020); https://doi.org/10.1145/3334480.3383092
21. Soulmachines Inc. https://www.soulmachines.com/2018/08/meet-will-vectors-new-renewable-energy-educator-in-schools/ (accessed 25 April 2021).
22. Deci, E. L., Vallerand, R. J., Pelletier, L. G. & Ryan, R. M. Motivation and education: the self-determination perspective. *Educational Psychol.* **26**, 325–346 (1991).
23. Skinner, E. A. & Belmont, M. J. Motivation in the classroom: reciprocal effects of teacher behavior and student engagement across the school year. *J. Educ. Psychol.* **85**, 571–581 (1993).
24. Alavi, M. & Leidner, D. E. Research commentary: technology-mediated learning—a call for greater depth and breadth of research. *Inf. Syst. Res.* **12**, 1–10 (2001).
25. Hudson, I. & Hurter, J. Avatar types matter: review of avatar literature for performance purposes. In *Proc. International Conference on Virtual, Augmented and Mixed Reality* 14–21 (Springer, 2016).
26. Kosmyna, N., Gross, A. & Maes, P. 'The thinking cap 2.0' preliminary study on fostering growth mindset of children by means of electroencephalography and perceived magic using artifacts from fictional sci-fi universes. In *Proc. Interaction Design and Children Conference* 458–469 (ACM, 2020).
27. Edwards, C. Male professor turns himself into anime schoolgirl to teach students remotely during coronavirus lockdown. *The U.S. Sun* (18 March 2020); https://www.the-sun.com/lifestyle/tech/556889/male-professor-turns-himself-into-anime-schoolgirl-to-teach-students-remotely-during-coronavirus-lockdown/
28. Kilteni, K., Bergstrom, I. & Slater, M. Drumming in immersive virtual reality: the body shapes the way we play. *IEEE Trans. Vis. Comput. Graph.* **19**, 597–605 (2013).
29. Peck, T. C., Good, J. J. & Bourne, K. A. Inducing and mitigating stereotype threat through gendered virtual body-swap illusions. In *Proc. 2020 CHI Conference on Human Factors in Computing Systems* 1–13 (ACM, 2020).
30. Guegan, J., Buisine, S., Mantelet, F., Maranzana, N. & Segonds, F. Avatar-mediated creativity: when embodying inventors makes engineers more creative. *Comput. Human Behav.* **61**, 165–175 (2016).

31. Banakou, D., Kishore, S. & Slater, M. Virtually being Einstein results in an improvement in cognitive task performance and a decrease in age bias. *Front. Psychol.* **9**, 917 (2018).
32. Leong, J. et al. Exploring the use of real-time camera filters on embodiment and creativity. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* 1–7 (ACM, 2021).
33. Ames, C. Motivation: what teachers need to know. *Teach. Coll. Rec.* **91**, 409–421 (1990).
34. Henderlong, J. & Lepper, M. R. The effects of praise on children's intrinsic motivation: a review and synthesis. *Psychol. Bull.* **128**, 774–795 (2002).
35. Bandura, A. Self-efficacy: toward a unifying theory of behavioral change. *Psychol. Rev.* **84**, 191–215 (1977).
36. Shiomi, M., Okumura, S., Kimoto, M., Iio, T. & Shimohara, K. Two is better than one: social rewards from two agents enhance offline improvements in motor skills more than single agent. *PLoS ONE* **15**, e0240622 (2020).
37. Kory-Westlund, J. M. & Breazeal, C. A long-term study of young children's rapport, social emulation and language learning with a peer-like robot playmate in preschool. *Front. Robot. AI* **6**, 81 (2019).
38. Musek, J. & Polic, M. in *Encyclopedia of Quality of Life and Well-Being Research* (ed. Michalos, A. C.) 4752–4755 (Springer, 2014).
39. National Institute of Mental Health. Mental illness; https://www.nimh.nih.gov/health/statistics/ (accessed 8 October 2021).
40. Torous, J. & Weiss Roberts, L. Needed innovation in digital health and smartphone applications for mental health transparency and trusts. *JAMA Psychiatry.* **74**, 437–438 (2017).
41. Zhou, L., Gao, J., Li, D. & Shum, H.-Y. The design and implementation of Xiaoice, an empathetic social chatbot. *Comput. Linguistics* **46**, 53–93 (2020).
42. Laranjo, L. et al. Conversational agents in healthcare: a systematic review. *J. Am. Med. Inform. Assoc.* **25**, 1248–1258 (2018).
43. Gaffney, H., Mansell, W. & Tai, S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. *JMIR Mental Health* **6**, e14166 (2019).
44. Loveys, K., Fricchione, G., Kolappa, K., Sagar, M. & Broadbent, E. Reducing patient loneliness with artificial agents: design insights from evolutionary neuropsychiatry. *J. Med. Internet Res.* **21**, e13664 (2019).
45. Johnson, L. A. & Caldwell, B. E. Race, gender and therapist confidence: effects on satisfaction with the therapeutic relationship in MFT. *Am. J. Family Therapy* **39**, 307–324 (2011).
46. Banerjee, A. & Sanyal, D. Dynamics of doctor-patient relationship: a cross-sectional study on concordance, trust and patient enablement. *J. Family Community Med.* **19**, 12–19 (2012).
47. Matsangidou, M. et al. 'Now i can see me' designing a multi-user virtual reality remote psychotherapy for body weight and shape concerns. *Hum. Comput. Interact.* https://doi.org/10.1080/07370024.2020.1788945 (2020).
48. Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S. & Torous, J. B. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *Can. J. Psychiatry* **64**, 456–464 (2019).
49. Carvalho, M. R. D., Freire, R. C. & Nardi, A. E. Virtual reality as a mechanism for exposure therapy. *World J. Biol. Psychiatry* **11**, 220–230 (2010).
50. Hershfield, H. E. et al. Increasing saving behavior through age-progressed renderings of the future self. *J. Mark. Res.* **48**, S23–S37 (2011).
51. Son, G.-R., Therrien, B. & Whall, A. Implicit memory and familiarity among elders with dementia. *J. Nurs. Scholarsh.* **34**, 263–267 (2002).
52. Pagnini, F. et al. Ageing as a mindset: a study protocol to rejuvenate older adults with a counterclockwise psychological intervention. *BMJ Open* **9**, e030411 (2019).
53. Massimi, M., Odom, W., Kirk, D. & Banks, R. HCI at the end of life: understanding death, dying and the digital. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* 4477–4480 (ACM, 2010).
54. Ohlheiser, A. The lonely reality of Zoom funerals. *MIT Technology Review* (13 April 2020); https://www.technologyreview.com/2020/04/13/999348/covid-19-grief-zoom-funerals/
55. Massimi, M. & Baecker, R. M. Dealing with death in design: developing systems for the bereaved. In *Proc. SIGCHI Conference on Human Factors in Computing Systems* 1001–1010 (ACM, 2011).
56. Newton, C. Speak, memory. *The Verge* (2016).
57. Hamilton, I. What is wisdom? *Business Insider* (17 November 2018); https://www.businessinsider.com/eternime-and-replika-giving-life-to-the-dead-with-new-technology-2018-11
58. Hayden, S. Mother meets recreation of her deceased child in VR https://www.roadtovr.com/mother-meets-recreation-of-deceased-child-in-vr/ (accessed 24 January 2021).
59. Brooker, C. & Harris, O. Be right back. Episode of *Black Mirror* (2013).
60. Villaronga, E. F. in *Emotional Design in Human–Robot Interaction* (eds Ayanoğlu, H. & Duarte, E.) 93–110 (Springer, 2019).
61. Kapur, A. et al. Non-invasive silent speech recognition in multiple sclerosis with dysphonia. In *Proc. Machine Learning for Health Workshop* 25–38 (PMLR, 2020).
62. Segalov, M. 'i choose to thrive': the man fighting motor neurone disease with cyborg technology. *The Guardian* (16 August 2021); https://www.theguardian.com/society/2020/aug/16/i-choose-to-thrive-the-man-fighting-motor-neurone-disease-with-cyborg-technology
63. Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A. & Black, M. J. Capture, learning and synthesis of 3D speaking styles. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) 10093–10103 (IEEE, 2019).
64. Li, T., Bolkart, T., Black, M. J., Li, H. & Romero, J. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* **36**, 194 (2017).
65. Vougioukas, K., Petridis, S. & Pantic, M. Realistic speech-driven facial animation with GANs. *Int. J. Comput. Vis.* **128**, 1398–1413 (2020).
66. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E. & Sebe, N. First order motion model for image animation. In *Advances in Neural Information Processing Systems* 32 (eds. Wallach, H. et al.) 7137–7147 (Curran Associates, 2019).
67. Jemine, C. et al. *Automatic Multispeaker Voice Cloning*. MSc thesis, Université de Liège (2019).
68. Sutton, R. E. Image manipulation: then and now. In Selected Readings from the Symposium of the International Visual Literacy Association (1993).
69. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. Preprint at https://arxiv.org/abs/1412.6572 (2014).
70. Rooney, B. Women and children first: technology and moral panic. *The Wall Street Journal* (11 June 2011); https://www.wsj.com/articles/BL-TEB-2814
71. Thompson, C. Texting isn't the first new technology thought to impair social skills. *Smithsonian Magazine* (March 20116); https://www.smithsonianmag.com/innovation/texting-isnt-first-new-technology-thought-impair-social-skills-180958091/
72. Müller, V. C. in *The Stanford Encyclopedia of Philosophy* summer 2021 edn (ed. Zalta, E. N.) (Stanford Univ., 2021); https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/
73. Langlois, S. 'Donald Trump' explains money laundering to his son-in-law in 'deepfake' video. *Market Watch* (19 September 2019); https://www.marketwatch.com/story/donald-trump-explains-money-laundering-to-his-son-in-law-in-deepfake-video-2019-09-18
74. Rajendra-Nicolucci, C. Language-generating A.I. is a free speech nightmare. *Slate* (30 September 2020); https://slate.com/technology/2020/09/language-ai-gpt-3-free-speech-harassment.html
75. Meskys, E., Kalpokiene, J., Jurcys, P. & Liaudanskas, A. Regulating deep fakes: legal and ethical considerations.*J. Intellect. Prop. Law Pract*. **15**, 24–31 (2019).
76. Wagner, T. L. & Blewer, A. 'The word real is no longer real': deepfakes, gender, and the challenges of AI-altered video. *Open Inf. Sci.* **3**, 32–46 (2019).
77. Fallis, D. The epistemic threat of deepfakes. *Philos. Technol.* https://doi.org/10.1007/s13347-020-00419-2 (2020).
78. In event of moon disaster (MIT Center For Advanced Virtuality, 2020); https://moondisaster.org
79. Ryan, E. The intersection of the Disney princess phenomenon and eating disorders. *Response The Journal of Popular and American Culture* https://responsejournal.net/issue/2016-08/article/intersection-disney-princess-phenomenon-and-eating-disorders (2016).
80. Burroughs, B. Youtube kids: the app economy and mobile parenting. *Soc. Media Soc.* https://doi.org/10.1177/2056305117707189 (2017).
81. Primack, B. A. et al. Social media use and perceived social isolation among young adults in the US. *Am. J. Prev. Med.* **53**, 1–8 (2017).
82. Umberson, D. & Karas Montez, J. Social relationships and health: a flashpoint for health policy. *J. Health Soc. Behav.* **51**, S54–S66 (2010).
83. Radesky, J. S. & Christakis, D. A. Increased screen time: implications for early childhood development and behavior. *Pediatr. Clin.* **63**, 827–839 (2016).
84. Shin, H. & Ryan, A. M. Friend influence on early adolescent disruptive behavior in the classroom: teacher emotional support matters. *Dev. Psychol.* **53**, 114–125 (2017).
85. Chiu, T. K. Student engagement in K-12 online learning amid COVID-19: a qualitative approach from a self-determination theory perspective. *Interactive Learn. Environ.* https://doi.org/10.1080/10494820.2021.1926289 (2021).
86. Abram, C. The most urgent threat of deepfakes isn't politics. It's porn. *Vox* (8 June 2020); https://www.vox.com/2020/6/8/21284005/urgent-threat-deepfakes-politics-porn-kristen-bell
87. Ankel, S. Many revenge porn victims consider suicide—why aren't schools doing more to stop it. *The Guardian* (7 May 2018).
88. Kugler, M. B. & Pace, C. Deepfake privacy: attitudes and regulation. Northwestern Public Law Research Paper, SSRN 21-04 (2021); https://ssrn.com/abstract=3781968
89. Malaria Must Die. David Beckham launches the world's first voice petition to end malaria https://malariamustdie.com/news/david-beckham-launches-worlds-first-voice-petition-end-malaria (accessed 15 July 2015).

90. Eshraghian, J. K. Human ownership of artificial creativity. *Nat. Mach. Intell.* **2**, 157–160 (2020).
91. Baek, S. Free Speech in the Digital Age: Deepfakes and the Marketplace of Ideas. Honors theses (PPE), University of Pennsylvania. Penn Libraries (2020).
92. Fazio, L. K., Brashier, N. M., Payne, B. K. & Marsh, E. J. Knowledge does not protect against illusory truth. *J. Exp. Psychol. Gen.* **144**, 993–1002 (2015).
93. Amershi, S. et al. Guidelines for human-AI interaction. In *Proc. 2019 CHI Conference on Human Factors in Computing Systems* **3**, 1–13 (ACM, 2019).
94. Naruniec, J., Helminger, L., Schroers, C. & Weber, R. M. High-resolution neural face swapping for visual effects. In *Proc. Computer Graphics Forum* Vol. 39.4, 173–184 (Wiley, 2020).
95. Ramanathan, M., Mishra, N. & Thalmann, N. M. Nadine humanoid social robotics platform. In *Proc. Computer Graphics International Conference* 490–496 (Springer, 2019).
96. Claes, P. et al. Modeling 3D facial shape from DNA. *PLoS Genet*. **10**, e1004224 (2014).
97. Zhu, W., Fan, X. & Zhang, Y. Applications and research trends of digital human models in the manufacturing industry. *Virtual Reality Intell. Hardware* **1**, 558–579 (2019).
98. Our first artificial human. Samsung Neon https://www.neon.life/ (accessed 10 July 2020).
99. SoulMachines Inc. Baby X: soul machines. SoulMachines Inc. https://www.soulmachines.com/ (accessed 10 July 2020).
100. Corentin, J. Real-Time Voice Cloning. MSc thesis, Université de Liège (2019).

Author contributions

P. Pataranutaporn developed the pipeline, assisted by V.D. The literature review was conducted by P. Pataranutaporn, V.D., J.L., P. Punpongsanon and M.S., who also contributed to the writing and editing of the manuscript. All other authors reviewed the manuscript. P. Pataranutaporn designed the figures. The pipeline was tested by D.N. The work was supervised by P.M. and M.S.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Misha Sra.