From low invasiveness to high control: How artificial intelligence allows to generate a large pool of standardized corpora at a lesser cost

Emmanuelle Kleinlogel, Laetitia Renier, Marianne Schmid Mast, Dinesh Babu Jayagopi, Kumar Shubham

TYPE Methods
PUBLISHED 09 May 2023
DOI 10.3389/fcomp.2023.1069352

OPEN ACCESS

EDITED BY
Mathieu Chollet,
University of Glasgow, United Kingdom

REVIEWED BY
Su Lei,
University of Southern California, United States
Tanaya Guha,
University of Glasgow, United Kingdom

*CORRESPONDENCE
Emmanuelle P. Kleinlogel
✉ emmanuelle.kleinlogel@univ-reunion.fr

RECEIVED 13 October 2022
ACCEPTED 13 April 2023
PUBLISHED 09 May 2023

CITATION
Kleinlogel EP, Renier LA, Schmid Mast M, Jayagopi DB and Shubham K (2023) From low invasiveness to high control: how artificial intelligence allows to generate a large pool of standardized corpora at a lesser cost. *Front. Comput. Sci.* 5:1069352. doi: 10.3389/fcomp.2023.1069352

From low invasiveness to high control: how artificial intelligence allows to generate a large pool of standardized corpora at a lesser cost

Emmanuelle P. Kleinlogel[1]*, Laetitia A. Renier[2], Marianne Schmid Mast[2], Dinesh Babu Jayagopi[3] and Kumar Shubham[4]

[1]CEMOI Laboratory, IAE Reunion, University of Reunion Island, Saint-Denis, France, [2]Faculty of Business and Economics, University of Lausanne, Lausanne, Switzerland, [3]International Institute of Information Technology, Bangalore, India, [4]Indian Institute of Science, Bangalore, India

The use of corpora represents a widespread methodology in interpersonal perception and impression formation studies. Nonetheless, the development of a corpus using the traditional approach involves a procedure that is both time- and cost-intensive and might lead to methodological flaws (e.g., high invasiveness). This might in turn lower the internal and external validities of the studies. Drawing on the technological advances in artificial intelligence and machine learning, we propose an innovative approach based on deepfake technology to develop corpora while tackling the challenges of the traditional approach. This technology makes it possible to generate synthetic videos showing individuals doing things that they have never done. Through an automatized process, this approach allows to create a large scale corpus at a lesser cost and in a short time frame. This method is characterized by a low degree of invasiveness given that it requires minimal input from participants (i.e., a single image or a short video) to generate a synthetic video of a person. Furthermore, this method allows a high degree of control over the content of the videos. As a first step, a referent video is created in which an actor performs the desired behavior. Then, based on this referent video and participant input, the videos that will compose the corpus are generated by a specific class of machine learning algorithms such that either the facial features or the behavior exhibited in the referent video are transposed to the face or the body of another person. In the present paper, we apply deepfake technology to the field of social skills and more specifically to interpersonal perception and impression formation studies and provide technical information to researchers who are interested in developing a corpus using this innovative technology.

KEYWORDS

corpus, interpersonal perception, impression formation, artificial intelligence, machine learning, deepfake, synthetic video

Being interviewed, participating in a professional meeting, making a presentation in front of an audience, or socializing with new colleagues are all situations that require social skills. In these situations, individuals typically want to convey a good impression. Investigating factors contributing to conveying a good impression is crucial to give clear recommendations to individuals seeking to improve how they are perceived by others. For instance, researchers might be interested in investigating the role of smiling or nodding (presence vs. absence of each behavior) during a job interview on the interviewee hireability. The traditional research approach usually investigates the effects of smiling and nodding on hireability by relying on naturalistic data through the creation of corpora. The creation of a corpus consists in compiling videos showing spontaneous or staged behaviors in a given social interaction to test a research question. These videos are then watched and evaluated by participants or judges to test the research question and hence to draw conclusions.

While the use of corpora allows investigating a wide range of research questions, this method presents some limitations. For instance, to investigate the role of smiling and nodding on hireability, researchers first need to create a corpus by recording individuals participating in a job interview. The goal is to collect videos showing variation in terms of smiling and nodding between targets. Targets are either participants or actors. As a next step, the videos are rated by a pool of judges to collect data on the perceived hireability of the individuals. The challenge for researchers in such studies is to disentangle the effects of smiling and nodding to assess the extent to which smiling on one hand and nodding on the other hand contribute to perceived hireability. Researchers might also be interested in assessing the interaction effect between smiling and nodding. Disentangling these two nonverbal behaviors can be done through coding. Nonetheless, this disentangling is not an easy task in natural stimuli given that they might be confounded in one person.

To tackle this challenge of isolating behaviors, researchers can decide to rely on actors over participants to create a corpus. The use of actors allows developing a corpus composed of staged behaviors, as opposed to spontaneous behaviors. For instance, an actor can be trained to either only smile or nod as well as to do both or neither, during the job interview. As a result, four videos compose the corpus in which the effects of smiling and nodding are clearly disentangled.

Nonetheless, this method presents several limitations. First, it is resource-intensive given that the actor should be trained to exhibit the desired behaviors. While participants are not trained, this method is also resource-intensive given that the corpus should be composed of a large pool of participants displaying a wide range of spontaneous behaviors. Second, researchers should also make sure that all staged behaviors are maintained constant across the videos to avoid methodological flaws such as the presence of alternative explanations that would have negative consequences in terms of causal claims (Shadish et al., 2002). Finally, the corpus should be composed of different targets to establish the external validity of the study, hence rendering the development of the corpus even more challenging (Shadish et al., 2002).

This paper proposes an innovative approach based on deepfake technology to develop a corpus that tackles the challenges of the traditional approach. This technology makes it possible to generate realistic videos that use artificial intelligence and machine learning techniques to portray people doing things that they have never done (Westerlund, 2019) or being someone different (e.g., younger/older self or feminine/masculine self; Shen et al., 2020; Zhu et al., 2020; Alaluf et al., 2021). Specifically, this technology makes it possible to combine the advantages of relying on participants (i.e., large sample size, diverse population) and on actors (i.e., control over the content of the videos). For instance, it makes it possible to generate a large number of videos showing different individuals either smiling, nodding, or both during a job interview, exactly in the same way, at the same time, and for the same duration, at a lower cost, and in a short time frame. It also allows creating another corpus where the same individuals have a different appearance such that they can appear older or more heavyset. Here, the behavior remains the same but individuals' features are changed.

The present paper contributes to the literature on interpersonal perception and impression formation in several ways. First, we contribute to the literature by presenting how this innovative approach tackles the challenges of the traditional approach. Second, we provide technical information to researchers who are interested in developing a corpus using these innovative technologies.

Innovative approach of generating synthetic videos

Overview of the innovative approach

A new class of algorithms has emerged that facilitates the creation of videos in which individual's behaviors or features are controlled or manipulated. These videos, popularly known as deepfakes, allow researchers to control the facial expression (Thies et al., 2016), head motion (Chen et al., 2020), lip sync (Prajwal et al., 2020), and body posture (Chan et al., 2019) of an individual in a given video. It is also possible to face swap (Nirkin et al., 2019; Xu et al., 2022) or to change the facial and/or body features of individuals (Shen et al., 2020; Zhu et al., 2020; Alaluf et al., 2021; Frühstück et al., 2022). Hence, the particularity of this technology is that it makes it possible to generate artificial videos in which the behavior and/or appearance of the individuals are purely fictional (Westerlund, 2019). This approach is particularly attractive to researchers willing to develop large-scale corpora in which they need to have a high degree of control over the content of the videos. Furthermore, this technology appears attractive because it is more accessible than we think and less resource-intensive than the traditional approach. To artificially generate a video portraying a specific individual, researchers need a target input (i.e., an image, such as a selfie, a full body image, or a short video of the target) and a referent video (i.e., a video of an actor performing the desired set of behaviors which is later transposed to the target's face and/or body).

Technical characteristics of deepfake technology

A specific class of machine learning algorithm known as generative adversarial network (GAN) (Goodfellow et al., 2020) generates the videos. The GAN is renowned for its capability to generate photorealistic images (Karras et al., 2019) and videos (Clark et al., 2019) that control the individual's appearance and motions (see also Bregler et al., 1997; Thies et al., 2016; Suwajanakorn et al., 2017). GAN takes inspiration from game theory to generate artificial multimedia output. During the training of such an algorithm, two models are trained simultaneously: a generator which tries to generate artificial videos or images while maintaining the identity of a given target, but with basic controls on the motion and the behavior exhibited by the target; and a discriminator which tries to discriminate the generated output from the images or real videos (short video of the individuals). The overall objective of the generator is to fool the discriminator by generating outputs that are indistinguishable from the real video while the discriminator tries to correctly distinguish the synthetic output from the real videos. As the training proceeds, it reaches an equilibrium state where the discriminator cannot distinguish artificial output from real world output and at the same time the generator cannot generate better images or videos.

The quality of the synthetic videos generated using GAN depends on multiple factors like the choice of loss functions, the network architecture, and the quality of the videos used in the training of these models. Loss functions are mathematical formulations of the training process and the desired result that one expects the GAN to produce. For example, the game theory-based objective discussed in the previous paragraph is known as the adversarial loss (Goodfellow et al., 2020). Similarly, loss functions like perceptual loss (Johnson et al., 2016) ensure that the synthetic videos are perceptually similar to the real videos for the human eye. Generally, deepfake methods use different types of loss functions to control the quality of the generated videos. Apart from the choice of loss functions, another important decision in GAN training is the choice of the network architecture for the generator. The GAN architecture plays a major role in determining the resolution of the video and in controlling the fine-grained details like the skin and hair texture in the genereated images. Two of the most famous generator architectures used in GAN for deepfake video generation are U-net (Ronneberger et al., 2015) and Pix2Pix-HD (Wang et al., 2018b). Similarly, styleGAN (Karras et al., 2019) is commonly used for gender and age based modifications.

Different types of synthetic video generation techniques exist depending on the desired behavior that needs to be transposed and the data required to train these GAN models. Specifically, the overall video generation process can be categorized into two techniques, namely face transfer and pose transfer techniques, primarily focusing on facial and body behaviors, respectively. On the one hand, the face transfer technique relies on face morphing (changing the face of an actor in a given video to the face of the intended individual; Natsume et al., 2018; Nirkin et al., 2019) and facial re-enactment (transferring the behavior of the actor to the upper body of the individual; Thies et al., 2016; Wu et al., 2018) to generate videos with upper body motions like head nodding, facial expression changes, and lip sync. On the other hand, the pose transfer technique allows researchers to transfer the complete body behavior of an actor, involving limbs and torso movement, onto the body of the target. A famous application of the pose transfer technique is "Everybody dance now" by Chan et al. (2019). Using just a few clips of the random motions of a given target, a new video is created in which that same target dances like an expert.

The vid2vid (Wang et al., 2018a) model released by Nvidia is one of the state-of-the-art models available online for users to generate these pose transfer videos. Similarly, based on the type of data required to train these GAN models to generate deepfake videos, further categorization is done over the available algorithms. For example, deepfake techniques traditionally require multiple videos of a target to make high-quality deepfake videos (Kietzmann et al., 2020), while new methods like one-shot deepfake algorithms can generate high-quality videos with just a single image of a target. First order motion model (FOMM; Siarohin et al., 2019) is a well-known one-shot deepfake generation model that can generate deepfake videos with the upper body motion of an actor using one single image of a target as the input.

Developing corpora: a two-step process

Applying this technology to corpus development, the video generation consists of a two-step process. First, researchers need to create a referent video (i.e., referent input) in which an actor performs the desired behavior. Second, input from participants (i.e., target input) is needed; either a standard portrait-like image (e.g., a selfie) or a standard motion video. Creating the standard motion video, for pose transfer, involves recording a video of each target performing a predefined set of motions (e.g., by asking the participant to watch a video of an actor exhibiting these predefined motions and to reproduce the motions at the same time while being video-recorded). The content of the predefined set of motions depends on the research topic and the research topic then determines which technology to use to develop the study material. Face transfer technology is appropriate to generate a corpus for which the research question focuses on the upper body of an individual (e.g., facial expression, head motion), such as in a job interview setting in which the interviewees are requested to sit at a desk. Because of the focus on the upper body and head/face region, face transfer techniques can also be used with GAN face editing (Shen et al., 2020; Zhu et al., 2020; Alaluf et al., 2021) to create deepfake videos with age and gender based variation. For example, one could render the face of a male target more feminine or the opposite (Zhu et al., 2020). Pose transfer technology is appropriate when the focus is on the whole body (e.g., posture) and for which the facial expressions are of lesser importance, such as in a public speech delivered in front of a large audience. Using pose transfer technology to transpose facial expressions would lead to a suboptimal rendering. Given that this technology focuses on capturing motions associated with a complete body, it struggles to accurately capture the fine-grained motions associated with facial muscle movements. Accordingly, on the one hand, if the study concerns the upper body, the predefined set of motions involves a set of dynamic facial expressions and head movements based on the desired motions of interest. On the other hand, if the study concerns the whole body, the predefined set of motions involves body movements, such as hand gestures, moving sideways, and head positions.

The standard motion video is then used to train a GAN model to generate videos maintaining the identity of the given target, but showing a new and different set of behaviors. A separate set of machine learning algorithms is used to extract the information related to facial expressions (Baltrusaitis et al., 2018), head motions (Murphy-Chutorian and Trivedi, 2008; Baltrusaitis et al., 2018), and joint movements (Cao et al., 2017) from the referent video (i.e., set of motions of the actor to be reproduced). This information about the desired motion, along with a person-specific GAN model, is later used to generate the set of synthetic videos showing the target making the same movements as the actor in the referent video.

Many state-of-the-art pose transfer methods like vid2vid, "Everybody dance now" (Chan et al., 2019), and face transfer techniques follow these steps to generate deepfake videos. Generally, these GAN models are tuned toward the identity of a specific individual, and a separate model is required to be trained for every new individual. However, considering that the behavior of the actor is common for all individuals, the information about the desired motion can be reused. Contrary to the above step, a one-shot deepfake technique like FOMM (Siarohin et al., 2019) does not require the training of separate GAN models for different targets and uses its own standard technique to track the motions of the actor. Deepfake video generation using FOMM requires just the facial image of the targets and the referent video. Finally, depending on the algorithm used and the computing-power available, the experimental material can be created in less than a week.

Deepfake videos are typically created on computers equipped with GPUs. On a system with one V100 GPU card with 16 GB of GPU memory, 48 core CPU, and 225 GB RAM, it takes close to 10 h to train a single vid2vid model using a 3-min video clip of a person at 15 frames per second and a resolution size of 256 × 256. However, with multiple cards and by reusing unused GPU memory, multiple videos can be created in the same time frame. Contrary to vid2vid, one-shot deepfake techniques like FOMM can create a 3-min video clip on a similar machine in less than 5 min.

Contributions of this innovative approach

Applied to the field of interpersonal perception and impression formation, we believe that deepfake technology is an opportunity to create new corpora easily and in a more standardized manner. First, this innovative approach is particularly suited for research which requires having a high degree of control over video-based corpora. For instance, imagine that researchers want to assess the effect of ethnicity on the perception of charisma during a public speech. To do so, they first need to develop a corpus showing individuals of different ethnic backgrounds delivering a charismatic speech. When creating the corpus, researchers should avoid the presence of alternative explanations as much as possible (Shadish et al., 2002). Ideally, the speakers should exhibit exactly the same behavior, at the same time, and the length of the speech should be the same, such that the only variation between the videos is the ethnicity of the speakers. Approaching this level of standardization with a traditional approach might be easier to access with a small sample of targets (i.e., staged behaviors displayed by a few trained actors). Nonetheless, a corpus composed of only one or a few individuals of different ethnic groups lowers the generalization of the findings, given that the results can be attributed to the speakers themselves (e.g., depending on their appearance such as their attractiveness and femininity/masculinity; see Todorov et al., 2015; Antonakis and Eubanks, 2017) and not to the ethnicity of the speakers. Hence, in the quest for external validity (Shadish et al., 2002), the sample of targets should be large enough, such that the corpus would include people with various appearances and from different ethnic groups.

For instance, imagine that researchers are interested in assessing the effect of Hispanic/Asian ethnicity on the perception of nonverbal charismatic signaling during a public speech and they target a sample of 80 participants for each ethnic group delivering a charismatic speech. Applying the innovative approach to this research question implies that researchers first need to create a unique referent video showing an actor delivering a charismatic speech using nonverbal tactics. Second, they need a pool of 160 participants (i.e., 80 targets per ethnicity) each of whom is asked to provide either (a) a selfie if the focus is on charisma in the face or (b) a short head-to-toe video if the focus in on charisma in the body. In the case of charisma in the face, face transfer is then used to produce a corpus of 160 synthetic videos showing charisma in facial expressions each starring a different person. In the case of charisma in the body, pose transfer is then used to produce a corpus of 160 synthetic videos showing charisma in body movements each starring a different person. Through the referent video and target input, deepfake technology allows to create a set of synthetic videos showing the 160 targets delivering exactly the same nonverbally charismatic speech, meaning delivering a speech while having exactly the same nonverbal behavior, at the same time, and with the same amplitude. The large sample size obtained in this deepfake-based corpus provides a wide range of target look in each group, hence minimizing the threat of alternative explanations that may be caused, for instance, by the low attractiveness and asymmetrical appearance of certain speakers and how they are perceived (see Antonakis and Eubanks, 2017).

Second, this innovative approach is also attractive to researchers interested in developing materials involving an experimental manipulation, from the straightforward presence/absence of behaviors to subtle changes in motion. For instance, to date little is known about the effects of gazing, nodding, and smiling during a job interview, and particularly in a highly controlled setting (Renier et al., work in preparation). Manipulating the presence and absence of these motions across videos while controlling other factors in the videos, such as other facial expressions, with a sufficiently large sample to establish external validity (Shadish et al., 2002), is highly challenging. Researchers willing to adopt such an experimental approach are traditionally constrained to develop material using a single or a restricted pool of actors as it takes time to train them to be able to act out an acceptable level of standardized behavior while exhibiting the desired manipulated behavior.

Drawing on this innovative approach, Renier et al. (work in preparation) addressed this research question using face transfer technology. In their study, they investigated the effect of interviewee immediacy behaviors on interview ouctomes (i.e., impressions in terms of competence, warmth, and overall favorable impression) by conducting two within-subject experiments. The experiments had four conditions with different combinations of nonverbal behaviors selected based on the literature and constraints associated to deepfake. Following the two-step process mentioned earlier, the researchers first created a referent video for each condition. For the first condition, the referent video showed the actor gazing while occasionally nodding and smiling when listening to the interviewer question. For the second condition, the referent video showed the actor gazing while occasionally nodding. For the third condition, the referent video showed the actor only gazing. Finally, for the fourth condition, the referent video showed the actor neither gazing, nor nodding, nor smiling. For the second step, they gathered target inputs by asking participants to submit one image of themselves (i.e., a selfie-like image) through an online questionnaire. After conducting a quality check on the submitted target input, the researchers used a sample of 157 selfies, which were then used to generate the synthetic videos in the four conditions. In total, the corpus was composed of 628 synthetic videos, representing the motions of the four referent videos transposed onto the face of each of the 157 participants. In Study 1, participants were asked to assess the impressions they thought they conveyed based on their own synthetic videos. In Study 2, the researchers then collected data on other-perception relative to interview outcomes by asking a pool of 823 judges to watch four videos (one video per condition, each video showed a different individual) and to report their impression of the interviewees (i.e., interview outcomes mentioned above).

Overall, this technology allows researchers to conduct studies using a more fine-grained approach by disentangling the effects of specific behavior such as head motions (e.g., nodding, gazing) or body posture (e.g., open gesture, moving). Furthermore, studying a wide range of behaviors from facial expressions (e.g., surprise, anger) to head motions (e.g., nodding, gazing) as well as body posture (e.g., open gesture, leaning, moving), including the study of small variations of such behavior, such as the intensity of a smile or gaze frequency during a job interview, is now possible. Hence, it is now conceivable to study the effect of a single specific cue or behavior across a wide range of individuals. For instance, it would be interesting to investigate how a cue or behavior is perceived in a specific situation (e.g., job interview, public speech) depending on the age, gender, ethnicity, haircut or hair color, clothes, or attractiveness of the individuals, to only cite a few examples. Such research would require having only one single referent video and collecting pictures or short videos of individuals with the desired appearance or demographics.

Besides looking for individuals with different appearance or demographics, one might want to study the same individuals, but experimentally manipulate their appearance (i.e., artificially change their features). More particularly, one might want to generate thin-slice videos where targets nod and smile while varying the appearance of the targets to make them appear younger or older, or as female or male. To illustrate, another example of experimental manipulation is the use of deepfake to manipulate the targets' behaviors and features. Bekbergenova et al. (2023) aimed at testing the effect of charismatic signaling and gender on market potential. In their study, they first transformed pictures of 13 male targets into female targets and vice versa (12 female targets into male targets). Second, they used face transfer to synthesize videos of the targets shown alongside the (non-) charismatic entrepreneurial pitch participants rated. This research provides additional evidence that deepfake technologies can be used to avoid idiosyncratic differences. Such an approach enabled them to perfectly control the nonverbal behaviors expressed across 25 targets while maintaining constant individual factors across conditions (e.g., age, appearance, attractiveness).

Comparing traditional and innovative approaches

The traditional approach of developing corpora

Social skills can be defined as a set of behaviors that individuals learn to effectively communicate and interact with others (Gresham and Elliott, 1990; McClelland and Morrison, 2003; Lynch and Simpson, 2010; Gresham, 2016). For instance, social skills include the ability to collaborate with others, initiate a relationship, participate in a discussion, organize a group meeting, help others, and show empathy. Hence, social skills are critical in interactions with others on an everyday basis at school, work as well as in personal settings (Salzberg et al., 1986; Wentzel, 2009; Lynch and Simpson, 2010; Davies et al., 2015; Agran et al., 2016; Bessa et al., 2019).

Social skills research is a broad field. For instance, literature has extensively examined the role of social skills interventions on skills development for individuals, in general, as well as for individuals with disorders or additional needs (e.g., Scattone, 2007; Bohlander et al., 2012; Gresham, 2016; Olivares-Olivares et al., 2019; Soares et al., 2021). Research has also investigated social skills in the fields of sports (e.g.,Vidoni and Ward, 2009; Bessa et al., 2019; Irmansyah et al., 2020), success in the workplace (e.g., Salzberg et al., 1986; Phillips et al., 2014; Agran et al., 2016), clinical communication between patients and clinicians (e.g., Schmid Mast, 2007; Schmid Mast et al., 2008; Carrard et al., 2016), public speaking (e.g., Bauth et al., 2019; Kleinlogel et al., 2021), and leadership (e.g., Groves, 2005; Riggio and Reichard, 2008; Singh, 2013; Tur et al., 2022). Examples of social skills research, and more specifically research related to interpersonal perception and impression formation, are identifying verbal and nonverbal behavioral factors that positively influence the impression made on an audience when delivering a public speech or the perceived hireability of an applicant during a job interview. Identifying such behaviors is crucial in order to give clear recommendations to individuals seeking to improve the impression they convey to others.

From a methodological lens, investigating interpersonal perception and impression formation can require the use of corpora in which either staged or spontaneous interpersonal interactions are captured (e.g., Street and Buller, 1987; Riggio and Reichard, 2008; Vidoni and Ward, 2009; Chollet et al., 2013; Frauendorfer et al., 2014; Carrard et al., 2016; Olivares-Olivares et al., 2019). Researchers might decide either to rely on secondary data or to collect data by video-recording social situations. The former option consists in using already existing data, i.e. using a corpus that was developed for the purpose of other research. This option is less resource-intensive, but it has some limitations given that the researchers must deal with the constraints of the existing set of video-recordings. Therefore, they do not have control over its content, such as the composition of the sample, the type of social interaction, the length of the interaction, or the angle of the recording (Blanch-Hartigan et al., 2018). Hence, it might be difficult for researchers to find secondary data fitting their research question.

These constraints make the latter option of creating a corpus attractive as the researchers can design all the parameters of the video-recordings. Traditionally, the development of a corpus requires the recruitment of targets (i.e., actors or participants) who are then asked to place themselves in a specific situation involving a social interaction while being video-recorded. Although widely relied upon in research, this approach raises several challenges in terms of resources and methods. Table 1 reports the different phases of corpus development for each approach (i.e., traditional and innovative), from the recruitment to the recording phase, and the extent to which each method is resource-intensive (i.e., in terms of time and budget) using a scale ranging from low, moderate, to intensive. Table 1 also reports the implications of each method in terms of (a) invasiveness and control and (b) the internal and external validities of the research. Through this table, we aim to highlight the extent to which this innovative approach contributes to corpus development by alleviating the challenges and methodological flaws of the traditional approach. We call future research to further discuss the points raised in our work and to provide empirical evidence of the added value of the innovative approach over the traditional approach of creating corpora in the field of social skills studies.

Two approaches requiring different resources

A corpus can be created using the traditional approach through two main methods, namely by relying on trained actors enacting staged behaviors or on participants (in laboratories or in real-life settings) showing spontaneous behaviors. In the first step, the development of a corpus requires a recruitment phase (see Table 1). Recruiting the targets (i.e., actors or participants) has several implications. The recruitment of actors typically happens when the researchers plan to give clear directions on how to behave during the social interaction of interest. In this case, a training phase is needed for the actors to exhibit the desired behavior. Once the actors are trained, the video-recordings phase can start.

Researchers might decide to recruit only one actor or several actors. Creating a corpus using several actors exhibiting the same behavior contributes to establishing the external validity by having a more diverse population in the corpus (Shadish et al., 2002). Choosing to rely on several actors might thus appear to be the best option. Nonetheless, using actors might be costly because it implies an intensive preparation before starting the recording. This preparation includes creating the protocol that the actors should follow during the video-recordings and a training phase to make sure that the actors will exhibit the desired behavior. The more the actors, the more costly the preparation becomes. This is because it takes time and requires an additional budget prior to the recording phase. Moreover, the actors' behavior might not be exactly the same, hence lowering the internal validity of the research. Furthermore, researchers should define precisely which actors to recruit for their video-recording given the implications in terms of external validity.

Participants are recruited when researchers plan to capture spontaneous behaviors or when it is not clear which cues would compose a specific behavior. This method is also used when researchers seek variations found in the population rather than having only a couple of actors in the corpus. Contrary to actors, participants are instructed to behave as they would in real life. This method thus does not require a training phase, but the recording phase is longer given the larger sample size. Finally, researchers might also decide to record individuals in real-life settings. This method requires a prospecting phase as researchers need to find organizations willing to participate in their study. Furthermore, this method raises data privacy issues. For instance, research investigating physician-patient communication should find physicians that agree to be videotaped during their consultations and patients should give their consent (e.g., Street and Buller, 1987; Schmid Mast et al., 2008).

Furthermore, an additional cost might arise from the use of participants (i.e., spontaneous behaviors) in more or less scripted situations as compared to the use of actors (i.e., staged behaviors): assessing and preparing the video-recording for a specific study. For instance, for the purpose of a study investigating listening behaviors during a job interview, in the case of staged behaviors, actors should be trained to exhibit combinations of nonverbal behaviors in a predefined setting (e.g., smiling when listening to an interviewer). Then, researchers should check whether the desired behavior is displayed correctly in the video-recording. In the case of spontaneous behaviors, researchers should first video-record the participants taking part in a job interview in which a confederate follows a script and plays the role of an interviewer. As a second step, researchers need to watch the videos for all participants and cut them to create thin-slice videos only when the participants are listening to the interview question (without the warranty of capturing smiling as a listening behavior). In the case of spontaneous behaviors captured through videos recorded in real-life settings, without any script whatsoever, researchers also need to watch all the videos and search for snippets related to their research question.

Overall, the traditional approach takes time. Several weeks or months are usually needed to develop the materials before obtaining the final corpus and only then can the data collection start. Furthermore, recruiting targets is particularly costly. A budget for the monetary compensation of the study participants needs to be established. The cost of the recording equipment should also be added to the budget.

Comparatively, on the one side, the presented innovative approach requires the recruitment of only one actor to create the referent video. This video then serves as a model to generate the synthetic videos of the participants. On the other side, similar to the traditional approach relying on actors, the actor should be trained to exhibit the desired behavior in the referent video. In the second step, depending on the technology used to create the synthetic videos, either a single image or a short video (standard motion video) of the targets is needed. In the case of face transfer, the data collection (image) can take place either online, by sending clear instructions to participants on how to create their selfie or video,

TABLE 1 Phases and characteristics of corpus development.

	Traditional approach		Innovative approach
	Trained targets	Spontaneous targets	Synthetic videos
Corpus development phase			
(1) Recruitment	Low	Intensive	Moderate
(2) Training/preparation	Moderate	Low	Low to moderate
(3) Recording	Low	Intensive	Low to moderate
Methodological implications			
(1) Invasiveness	Low to moderate	High	Low to moderate
(2) Control	High	Low	High
(3) Internal validity	High	Depends on the coding phase	High
(4) External validity	Low	High	High

or during laboratory sessions supervised by experimenters. In the case of collecting data online, a quality control step should check that participants correctly follow the instructions (through a visual inspection of each input). In the case of pose transfer, the innovative approach is more resource-intensive than when recording actors because it requires additional inputs (e.g., participant videos) which are not required in the traditional approach (see Table 1). Apart from the actor's referent video, pose transfer requires that researchers video-record participants enacting a pre-scripted set of behaviors in the laboratory (i.e., longer time required in the lab and to assess the quality of the input).

Once participant inputs are collected, the corpus is generated automatically for each participant as compared to the video-recording phase of the traditional approach, which requires a full recording session of the social situation for each target (actor, participant, or individual in real-life settings). Overall, the innovative approach is less resource-intensive than the traditional approach of creating a corpus using participants or individuals in real-life settings.

Methodological implications of each approach

Invasiveness

While the resource-intensive challenges of the traditional approach can be overcome by having a well-organized schedule and appropriate funding, especially in the case of recording spontaneous behaviors, this approach presents several methodological challenges. First, video-recording individuals during a social task to capture spontaneous behaviors might be obtrusive and thus might lower the ecological aspect (i.e., naturalness and realism) of the study material. Whereas actors are expected to be more accustomed to being video-recorded, it is plausible that individuals, both in laboratory and in real-life settings, are intimidated by the presence of one or several cameras (e.g., Herzmark, 1985; Coleman, 2000). Low degrees of naturalness and realism prevent the generalization of the findings to other settings, hence lowering the external validity of the research (Shadish et al., 2002). The innovative approach is virtually unobtrusive because no camera invades the recorded social interaction (see Table 1).

Control over video content

Video-recording participants or individuals in real-life settings implies that each participant behaves spontaneously and hence behaves differently. As a consequence, researchers have low control over the content of the videos in the corpus (see Table 1). This constitutes an important methodological flaw in studies in which it is crucial to avoid variation (as much as possible) other than the experimental manipulation or the appearance of the individuals throughout the corpus.

The lack of control in the traditional approach can be overcome in part by relying on actors. Training targets allows to control the behavior featured in the video-recordings such that actors show the behavior learnt during the training (e.g., McGovern and Tinsley, 1978; Dovidio and Ellyson, 1982). Therefore, prior to the study researchers need to identify exactly which behavior they wish to target (i.e., which, when, and how the behaviors need to be expressed) and to create a protocol that the actors should follow during the video-recordings (e.g., McGovern and Tinsley, 1978; Teven, 2007; Lybarger et al., 2017). It is noteworthy that perfect control is not achievable. For instance, actors can be trained to smile with a specific intensity during a job interview, but in practice they can still show different intensity of smiling throughout the interview, which might subsequently influence their hireability scores rated by judges (Ruben et al., 2015). Another main challenge of relying on actors is related to the external validity of the study (Shadish et al., 2002). Using a single actor or a small sample of actors to develop a corpus implies that the findings may be attributed to the actor(s) and not to the studied behavior. For instance, if the actors are all men, can the findings be generalized to women as well? This question highlights the issue of actor gender, but other issues which prevent the generalization of the findings are also highly relevant (e.g., age, ethnicity, clothes, attractiveness). Relying on several actors leads to a higher external validity of the study, but it potentially decreases its internal validity given that having perfect control over the behavior of several actors is almost impossible, depending on the complexity of the behaviors under study and the length of the social interaction to record.

The added value of the innovative approach as compared to the video-recording of spontaneous behaviors is that researchers have a high degree of control over the behaviors displayed in the videos. Given that the referent video serves as a model to create the synthetic videos, no variation can be observed across the generated videos: all synthetic videos show exactly the same behavior (e.g., same frequency, same intensity), at the same time, and for the same duration. The only difference between the videos is related to the research question. For instance, if researchers are interested in investigating the effects of gender in a social interaction situation, the only difference between the videos would be the gender of the individuals. If researchers are interested in the effects of three different smile intensities during a persuasive speech, then three referent videos are needed in which an actor, in the shoes of a speaker, smiles with three different intensities during a public speech. If researchers are interested in the interaction of gender (female vs. male), culture (e.g., Swiss vs. Indian), and smile intensities on interview outcomes, the three same referent videos will be used to generate the experimental video set but the pool of target inputs will have to be updated so as to be composed of female and male participants from Switzerland and from India. Thus, this innovative approach also presents a crucial advantage over the use of actors in the traditional approach. Once the referent videos are created, an unlimited number of corpora can be created based on target inputs. For instance, related to the smile intensity study, the innovative approach can lead to the creation of a corpus composed of 150 videos in which the three different motions are transposed to the face of a sample of 50 targets (using their selfies as input). This corpus can then be used to conduct a between-subject or a within-subject experiment composed of a three-level factor (i.e., smile intensity). Relying on a set of 50 different individuals with various appearance, exhibiting each of the three desired behaviors contributes to the generalization of the findings. Reaching the same sample size through the type of targets used in the traditional method (actors or participants) would be highly resource-intensive.

Limitations, practical implications, and technological advances

Naturalness and realism of the corpus

One of the main challenges when generating synthetic videos is to create high-quality videos in terms of naturalness and realism. To discuss this challenge, we can compare synthetic videos to research tools generated using virtual reality technologies. Virtual reality technologies allow to create artificial environments and social situations designed to mimic real-life settings. In virtual reality-based videos or immersive scenarios, a high degree of naturalness and the realism of virtual humans is associated to the uncanny valley effect. The uncanny valley effect consists in perceiving the avatars (i.e., the virtual and artificial human) as eerie in the specific case of a high quality rendering because they become too humanlike without being human, and hence lead to negative feelings (see Mori et al., 2012). Applied to the generation of synthetic videos using the proposed innovative approach, it is rather unlikely that this effect holds given that the goal of creating synthetic videos is to rely on new technologies to develop high quality videos comparable to the generation of videos using the traditional approach. Contrary to avatars in virtual reality-based videos, individuals in synthetic videos are expected to have a perfect humanlike appearance and behavior. Hence, we expect the degree of naturalness and realism to be of particular importance with regards to synthetic videos such that a low degree might potentially negatively affect judge perception. First, poor quality videos might disturb the video evaluation process. While in the case of the uncanny valley effect judges tend to focus on seeking imperfection when watching high quality virtual reality-based videos (i.e., high realism), judges watching synthetic videos might also focus on imperfection but we expect it to happen in the specific case of poor quality videos (i.e., low realism). Second, poor quality videos might lower the perceived credibility of the research material and hence of the study, which might consciously or unconsciously bias the evaluation of judges as well as lower the extent to which they are involved in the study.

Despite technological advances in generating good quality output, there are still some major limitations to the generation of these synthetic videos. Many of these methods perform poorly in variable backgrounds and poor lighting. Furthermore, methods like pose transfer struggle to generate good quality output for fine-grained motions associated with fingers, lips, and facial muscle movements (Ivan et al., 2021). Similarly, deepfake techniques relying on a selfie can generate unnatural teeth and lip movement for any behavior which exhibits extreme expression. Additionally, for both technologies, overlapping limbs (crossed legs or hands in front of the face) can lead to artifacts in the generated videos.

The naturalness and realism of synthetic videos depend largely on the quality and quantity of the videos used to train the GAN model. While an ideal deepfake video creation requires multiple videos of an individual with different backgrounds and lighting variations to account for similar variations in the intended motion of an actor (Das et al., 2021), a high quality output can also be generated using a few minutes of video by standardizing the recording environment for both the referent and target inputs. One way to achieve this is by creating a laboratory setting where both the referent actor and the targets are recorded under the same conditions (e.g., with uniform lighting conditions, a plain background, using the same camera, and at a uniform distance from the recording device).

The similarity of the behaviors displayed by the actor in the referent video and by the targets in the standard motion videos also contributes to generate high quality synthetic videos. If the targets recruited to create the standard motion videos also take part in the actual study (by watching the generated synthetic videos), then the researchers should make sure that these targets cannot guess the research question addressed in the study while recording their input video (here a standard motion video) in order to avoid demand effects. As a solution, researchers can ask participants to perform the desired (or similar) motions to be studied as well as irrelevant movements (e.g., random dance moves). When the generated videos are used as demonstration videos for participants to train their nonverbal behaviors in subsequent tasks, mixing the studied behaviors with unrelated moves also allows avoiding potential motor learning.

Apart from the motions, the appearance of the targets and the actor also plays a key role in generating realistic output. In many of the deepfake-based experiments (Chan et al., 2019; Lyu, 2020), it has been shown that facial and body accessories like earrings, glasses, and loose clothes or clothes with multiple graphic designs affect the quality of the generated videos. Similarly, facial hair can generate artifacts. To avoid these issues, both the targets and the actor should avoid these accessories and wear plain clothes that fit well. To optimize the quality of the features extracted from the referent and target videos and thus the generated videos, the recording setting should also offer the greatest visual contrast from clothes to background.

The same limitations apply to deepfake techniques relying on selfies. The variation in the background, lighting, and appearance between the referent video and the target input (i.e., the selfies) shared by the participants can affect the quality of the generated output. To ensure optimal quality, images can be collected in a laboratory setting so that the researchers have control over various factors (e.g., background, lighting, camera). Prior to the laboratory session, participants should be informed about requirements such as wearing a simple dark or colored T-shirt without any pattern. Researchers should also make sure that the laboratory is appropriate for such image and video-recording sessions (e.g., white or light background). As a less resource-intensive alternative, researchers might also collect the images online. Given that targets can take their selfie and send it by email, this solution requires no specific location, material, human resource, or compensation budget from researchers to collect the inputs. Nonetheless, it increases the risk of poor quality inputs. In this case, detailed instructions should be provided to the targets regarding how to take an appropriate selfie. For example, targets can be asked to take selfies in a plain background and in good lighting conditions without any facial accessories. Appendix 1 reports an example of instructions to follow to generate high quality inputs. For studies requiring videos as target inputs, we strongly recommend collecting the videos in a laboratory setting to ensure high similarity between the referent and target videos in terms of background, framing, lighting, appearance, and camera resolution. In this case, prior to the laboratory session, targets should be informed that during the laboratory session they should not wear loose clothes, avoid wearing light/white clothes (if light background given that a high contrast is needed), have their hair tied back, and remove their glasses.

Practical implications

As a first practical implication, we recommend that researchers set up a coding procedure to ensure the quality of their newly developed corpus. We suggest a three-step coding procedure prior to starting the data collection of the actual study relying on said corpus. As a first step, researchers might want to ensure the quality of the inputs sent by participants (selfies of videos) by coding whether the inputs are in accordance with the standards set to generate good quality outputs as discussed earlier (e.g., background, lighting, clothes). The second step consists in coding the referent video(s) to ensure that the desired behavior is shown in accordance to the researchers' expectations. The third step consists in coding a randomly selected set of synthetic videos to ensure the quality of the desired behavior transfer as well as the overall quality of the synthetic videos. This final coding step is closely related to the first coding step given that not checking the quality of the target inputs or checking only a randomly selected set of selfies or videos might imply a high rate of poor quality outputs. In this case, researchers might decide to drop these low quality outputs from the corpus, which might require creating additional outputs to reach the desired number of videos in the corpus if a large pool of outputs are removed from the corpus. Oversampling the number of targets by recruiting additional people and hence, the quantity of target inputs, is recommended to avoid this pitfall. Appendix 2 reports an example of a three-step coding procedure that researchers can follow to ensure the quality of their newly developed corpus.

Second, it is noteworthy that the presented innovative approach requires expertise. Developing the algorithms and mastering the video generation process are the domain of computer scientists. Hence, researchers in the field of interpersonal perception and impression formation, in the field of (nonverbal) behavior studies, seeking to adopt this innovative approach would benefit from collaborating with computer scientists. Such multidisciplinary collaborations would also benefit computer scientists. Through a better understanding of the needs of researchers in this field, computer scientists cannot only endeavor to provide solutions for the diverse challenges of the traditional approach, but also develop tools that are more validly anchored in human interactions. Furthermore, apart from the costs associated with the corpus development phases discussed earlier, this approach requires additional costs given that it is computationally intensive to generate high quality synthetic videos. To illustrate, Table 2 reports an example of the costs associated with the creation of a corpus using the innovative approach and the traditional approach (targets as participants). Table 2 provides a cost estimation for a study similar to that of Renier et al. (work in preparation) on nonverbal immediacy behavior as presented earlier. Imagine that for this study the researchers target the development of a corpus composed of four experimental conditions and a sample of 180 participants. The same 180 participants will appear in each condition. Hence, this design implies the development of a corpus composed of 720 videos (180*4). Each video is expected to last 12 s. Table 2 reports a cost estimation for both approaches in each step of the corpus development, from the recruitment of participants in a pre-selection phase to the corpus video quality check.

To compute the costs, we rely on a fixed hourly wage of 27 Swiss Francs (about 29$) to compensate the targets, and to remunerate the research assistant and referent actor. Estimated time and costs in each step are based on the authors' previous experience and may vary depending on the complexity of the studied behavior, the deepfake technology used (e.g., face transfer, body transfer), previous experience related to the assigned task of the individuals involved in the project (e.g., research assistant, actor), and the hourly wage that might differ across countries. With regards to the

TABLE 2 Costs estimation example to create a corpus using the innovative approach and the traditional approach.

	Material development to create a corpus of 720 videos (180 targets declined into 4 conditions)			Material development to create a corpus of 720 videos (180 targets declined into 4 conditions)		
	Description	Duration (min)	Cost (CHF)	Description	Duration (min)	Cost (CHF)
Target input	Target compensation (preselection questionnaire and selfie submission by participants)	180*3	243.00	Target compensation (training, task, and 15 min. of participant video-recording)	180*15	1,215.00
	Filing and quality assessment of target input (i.e., to collect, assess, and archive submitted selfies)—Done by a research assistant	180*5	405.00	Filing and quality assessment of target input (i.e., to collect, assess, and archive recorded videos, to view and prepare videos to create thin-slices of 12 s displaying targeted nonverbal behavior)—Done by a research assistant	180*45	3,645.00
Referent inputs	Production of 4 referent videos (including actor training and pre-test of the technology and its output)—Done by a computer scientist collaborating on the project	60	27.00	–	–	–
Cloud and computing services	Swiss-based, Computing need to generate 180*4 deepfake videos System specification: 225 GB RAM, 48 CORES, 4 dedicated GPU NVIDIA TESLA V100, 500 GB local disk Production time depends on computer (greater GPU/RAM, faster production)	14,400	800.00	–	–	–
Quality assessment of outputs	Filing and quality assessment of outputs, i.e. to file all outputs and to perform a quality check on a random sample for 10 participants (4*10 generated videos)—Done by a research assistant	150	67.50	Quality assessment of outputs, i.e. to check the filing of and to perform a quality check on all outputs (4*180 thin-slice videos)—Done by a research assistant	150	243.00
Total			1,542.50			5,103.00

This cost estimation does not take into account the cost associated with the salary of the researchers and computer scientists involved in the study. The hourly wage used for the targets (participants), actors, and research assistants is fixed at CHF 27.00.

10.3389/fcomp.2023.1069352

cloud and computing services, previous experience taught us that we needed to use only Swiss-based cloud and computing services (as opposed to US-based solutions such as Amazon Web Services) due to European and Swiss data protection laws given that we were dealing with personal information. Relying on a Swiss-based cloud and computing services enables us, first, to respect data protection laws (all data remained in the EU and were treated in the EU even if the computer scientists were located in a data protection laws non-compliant country) and to rent computers rather than buying them for a specific projects (the computing service offers different levels of computing performance implying more or less expenses and more or less production time). Overall, this example shows that despite the fact that this innovative approach might involve additional steps in the corpus development (i.e., development of referent inputs, cloud and computing services related steps), it is still less cost-intensive (about 3–4 times less costly) than the traditional approach.

Third, the presented innovative approach contributes to corpora development by facilitating data sharing. Although not specifically addressed in this paper, the traditional approach also has issues related to data protection, hence adding another challenge to the progress of this research (see Renier et al., 2021). Laws related to data privacy require researchers to collect consent forms from participants to use their videos. Participants should be explicitly informed about the use (i.e., who will have access to the data?) and the storage of the videos (i.e., where and how long will the data be stored?). This limits data sharing possibilities given that researchers may not know in advance about other research opportunities which require sharing their corpus with other researchers. Furthermore, not explicitly collecting participant consent to share participant videos with other researchers makes it impossible to use secondary data.

Data sharing is less of an issue with regards to the innovative approach, given that only sharing the referent video is needed to create a large corpus pool. Therefore, once researchers have collected the data sharing approval consent form from the actor(s), the research material can be shared by storing the referent video(s) on an open data sharing platform without the data sharing concerns that researchers face with video-recordings from real participants. In this case, then researchers need to collect a new set of target inputs to create their corpus using the shared referent video. Additionally, attention should be paid to the use of cloud computing platforms in cases where the researchers do not have the adequate computing power to generate synthetic videos. For instance, for some countries even if the researchers obtain participant consent, it is not legally acceptable to use American-based platforms. The same goes for collaborating with teams of computer scientists from countries where the General Data Protection Regulation (GDPR) is not enforced. We strongly urge researchers to obtain information about the data protection laws observed by their institution and country. Obtaining clear information about cloud computing services and underlying laws is also of added value. For instance, all research groups do not possess the same high performance computing services, which directly affect the production time necessary for videos syntheses. Choosing a cloud computing service based in a GDPR-compliant country and paying for a higher performance machine can thus be more appropriate.

Finally, it is noteworthy that the democratization of such technologies might lead to wrong use and abuse if it is not monitored and supervised. Broadly speaking, given that these technologies allow swapping any face and body in a video, anybody can become the main character of an artificial video in which one can observe herself/himself or someone else appearing differently, or saying or doing something they did not intend to say or do. Such videos can damage an individual's personal and professional life. This warning is relevant given that technological advances allow to create high quality deepfake such that the majority of individuals cannot detect the synthetic feature of the videos (see Korshunov and Marcel, 2020). Therefore, the deepfake technology development is a double-edged sword. On the one side, greater realism is an advantage for research given that this technology offers high internal and external validity material. On the other side, it provides threats for society.

More specific to research, the development of synthetic videos might contribute to spread fake content and practices, that might negatively influence participants. For instance, synthetic videos might contribute to misinformation spread as well as the diffusion of stereotypes and/or wrong practices by swapping individuals' features, face, and/or behavior, depending on the research question addressed in the study (e.g., research questions investigating factors fostering the expression of stereotypes or the extent to which a leader is perceived as toxic by her/his subordinates). Such videos might also create psychological discomfort among individuals if for instance these recordings show negative emotionally arousing content (e.g., aggressive behaviors) or if they show someone familiar to the individuals doing something unethical. Research should reflect on how to control these new practices. To this purpose, we perceive the consent form to play a crucial role in informing individuals (participants, judges) of the nature of the videos and hence in controlling the usage of one's image as well as in avoiding deception and negative consequences that might be induced by psychological discomfort during the study. For instance, ethics committees might require participants to sign an informed consent explicitly stating the use of synthetic videos as well as notifying potential discomfort that might be caused. We redirect researchers interested in this topic to Westerlund's (2019) article on the benefits and threats of deepfake technology for society, as well on solutions that are available to combat the wrong use and abuse of this technology.

Upcoming technological advances

To date, current technologies allow to develop corpora for which the research question focuses on nonverbal behavior for video recordings in a controlled setting. However, in recent years, new research directions have emerged in the field of deepfake video creation. One notable work in this area has been the creation of audio-based deepfakes (Shen et al., 2018; Khanjani et al., 2021). With such a technology, researchers can create an audio-recording using the voice of a given person to utter any arbitrary sentence that the person has never actually uttered. Similar research is being carried out to improve lip-sync using few shot methods (Prajwal et al., 2020). Few shot methods are techniques that use less amount of data from participants to generate deepfake videos. For example, generating proper lip-sync using traditional techniques requires hours of recording participants. Few shot methods generate the same quality of outputs with fewer minutes of clip or with just a few images of targets. This audio-based deepfake, along with the proper lip-sync generation, can further open up possibilities for social skills experiments and studies by placing the focus on verbal behavior or two-sided social interaction where partners talk and act. Researchers are also focusing on tools to control the posture and body weight of an individual in an image (Frühstück et al., 2022). Such body editing GANs along with pose transfer techniques open up the possibility to explore the impact of different body weights in social situations. Furthermore, in the field of video-based deepfake generation, researchers are coming up with new robust techniques to generate these videos from online recordings (Zhou et al., 2019). Companies, like synthesia[1] have commercialized these deepfake generation technologies to create advertisement videos at a low cost.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

EK, LR, and MS conceived the presented idea, refined the main ideas and proof outline, and contributed to the final manuscript. All authors contributed different parts of writing to the manuscript with EK being in charge of coordinating and writing the most extensive part of the manuscript and discussed the initial content.

1 See for more information: https://www.synthesia.io/.

Funding

This work was supported by a grant from the Swiss National Science Foundation (CRSII5_183564) awarded to MS.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2023.1069352/full#supplementary-material

References

Agran, M., Hughes, C., Thoma, C. A., and Scott, L. A. (2016). Employment social skills: what skills are really valued? *Career Dev. Transit. Except. Individ.* 39, 111–120. doi: 10.1177/2165143414546741

Alaluf, Y., Patashnik, O., and Cohen-Or, D. (2021). Only a matter of style: age transformation using a style-based regression model. *ACM Trans. Graph.* 40, 1–12. doi: 10.1145/3450626.3459805

Antonakis, J., and Eubanks, D. L. (2017). Looking leadership in the face. *Curr. Dir. Psychol. Sci.* 26, 270–275. doi: 10.1177/0963721417705888

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). "Openface 2.0: facial behavior analysis toolkit," in *13th IEEE International Conference on Automatic Face and Gesture Recognition* (Xi'an: IEEE). doi: 10.1109/FG.2018.00019

Bauth, M. F., Angélico, A. P., and de Oliveira, D. C. R. (2019). Association between social skills, sociodemographic factors and self-statements during public speaking by university students. *Trends Psychol.* 27, 677–692. doi: 10.9788/TP2019.3-06

Bekbergenova, A., Schmid Mast, M., Antonakis, J., Krings, F., Renier, L. A., Shubham, K., et al. (2023). *Language in Entrepreneurial Pitching: Above and Beyond Gender Stereotypes* [Unpublished doctoral dissertation]. Lausanne: University of Lausanne.

Bessa, C., Hastie, P., Araújo, R., and Mesquita, I. (2019). What do we know about the development of personal and social skills within the sport education model: a systematic review. *J. Sci. Med. Sport* 18, 812–829.

Blanch-Hartigan, D., Ruben, M. A., Hall, J. A., and Schmid Mast, M. (2018). Measuring nonverbal behavior in clinical interactions: a pragmatic guide. *Patient Educ. Couns.* 101, 2209–2218. doi: 10.1016/j.pec.2018.08.013

Bohlander, A. J., Orlich, F., and Varley, C. K. (2012). Social skills training for children with autism. *Pediatr. Clin.* 59, 165–174. doi: 10.1016/j.pcl.2011.10.001

Bregler, C., Covell, M., and Slaney, M. (1997). "Video rewrite: driving visual speech with audio," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques* (Los Angeles, CA). doi: 10.1145/258734.258880

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. (2017). "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 7291–7299. doi: 10.1109/CVPR.2017.143

Carrard, V., Schmid Mast, M., and Cousin, G. (2016). Beyond "one size fits all": physician nonverbal adaptability to patients' need for paternalism and its positive consultation outcomes. *Health Commun.* 31, 1327–1333. doi: 10.1080/10410236.2015.1052871

Chan, C., Ginosar, S., Zhou, T., and Efros, A. A. (2019). "Everybody dance now," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul), 5933–5942. doi: 10.1109/ICCV.2019.00603

Chen, L., Cui, G., Liu, C., Li, Z., Kou, Z., Xu, Y., et al. (2020). "Talking-head generation with rhythmic head motion," in *European Conference on Computer Vision* (Glasgow), 35–51. doi: 10.1007/978-3-030-58545-7_3

Chollet, M., Ochs, M., and Pelachaud, C. (2013). "A multimodal corpus for the study of non-verbal behavior expressing interpersonal stances," in *IVA 2013 Workshop Multimodal Corpora: Beyond Audio and Video* (Edinburgh).

Clark, A., Donahue, J., and Simonyan, K. (2019). Adversarial video generation on complex datasets. *arXiv.* [preprint]. doi: 10.48550/arXiv.1907.06571

Coleman, T. (2000). Using video-recorded consultations for research in primary care: advantages and limitations. *Fam. Pract.* 17, 422–427. doi: 10.1093/fampra/17.5.422

Das, S., Seferbekov, S., Datta, A., Islam, M., and Amin, M. (2021). "Towards solving the deepfake problem: an analysis on improving deepfake detection using dynamic face augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal, QC). doi: 10.1109/ICCVW54120.2021.00421

Davies, M., Cooper, G., Kettler, R. J., and Elliott, S. N. (2015). Developing social skills of students with additional needs within the context of the Australian curriculum. *Australas. J. Spec. Educ.* 39, 37–55. doi: 10.1017/jse.2014.9

Dovidio, J. F., and Ellyson, S. L. (1982). Decoding visual dominance: attributions of power based on relative percentages of looking while speaking and looking while listening. *Soc. Psychol. Q.* 45, 106–113. doi: 10.2307/3033933

Frauendorfer, D., Schmid Mast, M., Nguyen, L., and Gatica-Perez, D. (2014). Nonverbal social sensing in action: unobtrusive recording and extracting of nonverbal behavior in social interactions illustrated with a research example. *J. Nonverbal Behav.* 38, 231–245. doi: 10.1007/s10919-014-0173-5

Frühstück, A., Singh, K. K., Shechtman, E., Mitra, N. J., Wonka, P., and Lu, J. (2022). "Insetgan for full-body image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA: IEEE), 7723–7732. doi: 10.1109/CVPR52688.2022.00757

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Commun. ACM* 63, 139–144. doi: 10.1145/3422622

Gresham, F. M. (2016). Social skills assessment and intervention for children and youth. *Camb. J. Educ.* 46, 319–332. doi: 10.1080/0305764X.2016.1195788

Gresham, F. M., and Elliott, S. N. (1990). *Social Skills Rating System: Manual.* Toronto, ON: Pearson Assessments. doi: 10.1037/t10269-000

Groves, K. S. (2005). Gender differences in social and emotional skills and charismatic leadership. *J. Lead. Organ. Stud.* 11, 30–46. doi: 10.1177/107179190501100303

Herzmark, G. (1985). Reactions of patients to video recording of consultations in general practice. *Br. Med. J.* 291, 315–317. doi: 10.1136/bmj.291.6491.315

Irmansyah, J., Lumintuarso, R., Sugiyanto, F., and Sukoco, P. (2020). Children's social skills through traditional sport games in primary schools. *Cakrawala Pendidik.* 39, 39–53. doi: 10.21831/cp.v39i1.28210

Ivan, V.-A., Mistreanu, I., Leica, A., Yoon, S.-J., Cheon, M., Lee, J., et al. (2021). "Improving key human features for pose transfer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, BC: IEEE). doi: 10.1109/ICCVW54120.2021.00223

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the Computer Vision–ECCV* (Cham), 694–711. doi: 10.1007/978-3-319-46475-6_43

Karras, T., Laine, S., and Aila, T. (2019). "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA), 4401–4410. doi: 10.1109/CVPR.2019.00453

Khanjani, Z., Watson, G., and Janeja, V. P. (2021). How deep are the fakes? Focusing on audio deepfake: a survey. *arXiv.* [preprint]. doi: 10.48550/arXiv.2111.14203

Kietzmann, J., Lee, L. W., McCarthy, I. P., and Kietzmann, T. C. (2020). Deepfakes: trick or treat? *Bus. Horiz.* 63, 135–146. doi: 10.1016/j.bushor.2019.11.006

Kleinlogel, E. P., Curdy, M., Rodrigues, J., Sandi, C., and Schmid Mast, M. (2021). Doppelganger-based training: imitating our virtual self to accelerate interpersonal skills learning. *PLoS ONE* 16, e0245960. doi: 10.1371/journal.pone.0245960

Korshunov, P., and Marcel, S. (2020). Deepfake detection: humans vs. machines. *arXiv*. [preprint]. doi: 10.48550/arXiv.2009.03155

Lybarger, J. E., Rancer, A. S., and Lin, Y. (2017). Superior–subordinate communication in the workplace: verbal aggression, nonverbal immediacy, and their joint effects on perceived superior credibility. *Commun. Res. Rep.* 34, 124–133. doi: 10.1080/08824096.2016.1252909

Lynch, S. A., and Simpson, C. G. (2010). Social skills: laying the foundation for success. *Dimens. Early Child.* 38, 3–12.

Lyu, S. (2020). "Deepfake detection: current challenges and next steps," *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)* (London: IEEE). doi: 10.1109/ICMEW46912.2020.9105991

McClelland, M. M., and Morrison, F. J. (2003). The emergence of learning-related social skills in preschool children. *Early Child. Res. Q.* 18, 206–224. doi: 10.1016/S0885-2006(03)00026-7

McGovern, T. V., and Tinsley, H. E. (1978). Interviewer evaluations of interviewee nonverbal behavior. *J. Vocat. Behav.* 13, 163–171. doi: 10.1016/0001-8791(78)90041-6

Mori, M., MacDorman, K., and Kageki, N. (2012). The uncanny valley. *IEEE Robot. Autom. Mag.* 19, 98–100. doi: 10.1109/MRA.2012.2192811

Murphy-Chutorian, E., and Trivedi, M. M. (2008). Head pose estimation in computer vision: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 607–626. doi: 10.1109/TPAMI.2008.106

Natsume, R., Yatagawa, T., and Morishima, S. (2018). "Fsnet: an identity-aware generative model for image-based face swapping," in *Asian Conference on Computer Vision* (Perth).

Nirkin, Y., Keller, Y., and Hassner, T. (2019). "Fsgan: subject agnostic face swapping and reenactment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul: IEEE), 7184–7193. doi: 10.1109/ICCV.2019.00728

Olivares-Olivares, P. J., Ortiz-González, P. F., and Olivares, J. (2019). Role of social skills training in adolescents with social anxiety disorder. *Int. J. Clin. Health Psychol.* 19, 41–48. doi: 10.1016/j.ijchp.2018.11.002

Phillips, B. N., Kaseroff, A. A., Fleming, A. R., and Huck, G. E. (2014). Work-related social skills: definitions and interventions in public vocational rehabilitation. *Rehabil. Psychol.* 59, 386. doi: 10.1037/rep0000011

Prajwal, K., Mukhopadhyay, R., Namboodiri, V. P., and Jawahar, C. (2020). "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia* (New York, NY). doi: 10.1145/3394171.3413532

Renier, L. A., Kleinlogel, E. P., Schmid Mast, M., Shubham, K., and Jayagopi, D. B. (work in preparation). Deepfake for the experimental study of nonverbal behaviors: Investigating perception of nonverbal immediacy behaviors using AI-generated characters.

Renier, L. A., Schmid Mast, M., Dael, N., and Kleinlogel, E. P. (2021). Nonverbal social sensing: what social sensing can and cannot do for the study of nonverbal behavior from video. *Front. Psychol.* 12, 606548. doi: 10.3389/fpsyg.2021.606548

Riggio, R. E., and Reichard, R. J. (2008). The emotional and social intelligences of effective leadership: an emotional and social skill approach. *J. Manag. Psychol.* 23, 169–185. doi: 10.1108/02683940810850808

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015* (Cham), 234–241. doi: 10.1007/978-3-319-24574-4_28

Ruben, M. A., Hall, J. A., and Schmid Mast, M. (2015). Smiling in a job interview: when less is more. *J. Soc. Psychol.* 155, 107–126. doi: 10.1080/00224545.2014.972312

Salzberg, C. L., Agran, M., and Lignugaris, B. (1986). Behaviors that contribute to entry-level employment a profile of five jobs. *Appl. Res. Ment. Retard.* 7, 299–314. doi: 10.1016/S0270-3092(86)80003-0

Scattone, D. (2007). Social skills interventions for children with autism. *Psychol. Sch.* 44, 717–726. doi: 10.1002/pits.20260

Schmid Mast, M. (2007). On the importance of nonverbal communication in the physician–patient interaction. *Patient Educ. Couns.* 67, 315–318. doi: 10.1016/j.pec.2007.03.005

Schmid Mast, M., Hall, J. A., Klöckner, C., and Choi, E. (2008). Physician gender affects how physician nonverbal behavior is related to patient satisfaction. *Med. Care* 46, 1212–1218. doi: 10.1097/MLR.0b013e31817e1877

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton: Mifflin and Company.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., et al. (2018). "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Calgary, AB). doi: 10.1109/ICASSP.2018.8461368

Shen, Y., Yang, C., Tang, X., and Zhou, B. (2020). Interfacegan: interpreting the disentangled face representation learned by gans. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 2004–2018. doi: 10.1109/TPAMI.2020.3034267

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., and Sebe, N. (2019). First order motion model for image animation. *Adv. Neural Inf. Process. Syst.* 32.

Singh, P. (2013). A collegial approach in understanding leadership as a social skill. *Int. Bus. Econ. Res. J.* 12, 489–502. doi: 10.19030/iber.v12i5.7824

Soares, E. E., Bausback, K., Beard, C. L., Higinbotham, M., Bunge, E. L., and Gengoux, G. W. (2021). Social skills training for autism spectrum disorder: a meta-analysis of in-person and technological interventions. *J. Technol. Behav. Sci.* 6, 166–180. doi: 10.1007/s41347-020-00177-0

Street, R. L., and Buller, D. B. (1987). Nonverbal response patterns in physician-patient interactions: a functional analysis. *J. Nonverbal Behav.* 11, 234–253. doi: 10.1007/BF00987255

Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: learning lip sync from audio. *ACM Trans. Graph.* 36, 1–13. doi: 10.1145/3072959.3073640

Teven, J. J. (2007). Effects of supervisor social influence, nonverbal immediacy, and biological sex on subordinates' perceptions of job satisfaction, liking, and supervisor credibility. *Commun. Q.* 55, 155–177. doi: 10.1080/01463370601036036

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). "Face2face: real-time face capture and reenactment of RGB videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 2387–2395. doi: 10.1145/2929464.2929475

Todorov, A., Olivola, C. Y., Dotsch, R., and Mende-Siedlecki, P. (2015). Social attributions from faces: determinants, consequences, accuracy, and functional significance. *Annu. Rev. Psychol.* 66, 519–545. doi: 10.1146/annurev-psych-113011-143831

Tur, B., Harstad, J., and Antonakis, J. (2022). Effect of charismatic signaling in social media settings: evidence from TED and Twitter. *Leadersh. Q.* 33, 101476. doi: 10.1016/j.leaqua.2020.101476

Vidoni, C., and Ward, P. (2009). Effects of fair play instruction on student social skills during a middle school sport education unit. *Phys. Educ. Sport Pedagogy* 14, 285–310. doi: 10.1080/17408980802225818

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., et al. (2018a). Video-to-video synthesis. *arXiv.* [preprint]. doi: 10.48550/arXiv.1808.06601

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018b). "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE: Salt Lake City, UT), 8798–8807. doi: 10.1109/CVPR.2018.00917

Wentzel, K. R. (2009). "Peers and academic functioning at school," in *Handbook of Peer Interactions, Relationships, and Groups*, eds K. H. Rubin, W. M. Bukowski, and B. Laursen (New York, NY: Guilford Press), 531–547.

Westerlund, M. (2019). The emergence of deepfake technology: a review. *Technol. Innov. Manag. Rev.* 9, 39–52. doi: 10.22215/timreview/1282

Wu, W., Zhang, Y., Li, C., Qian, C., and Loy, C. C. (2018). "ReenactGAN: learning to reenact faces via boundary transfer," *Proceedings of the European Conference on Computer Vision (ECCV)* (Cham), 603–619. doi: 10.1007/978-3-030-0124 6-5_37

Xu, Z., Hong, Z., Ding, C., Zhu, Z., Han, J., Liu, J., et al. (2022). Mobilefaceswap: a lightweight framework for video face swapping. *Proc. AAAI Conf. Artif. Intelli.* 36, 2973–2981. doi: 10.1609/aaai.v36i3.20203

Zhou, Y., Wang, Z., Fang, C., Bui, T., and Berg, T. (2019). "Dance dance generation: motion transfer for internet videos," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul: IEEE). doi: 10.1109/ICCVW.2019.00153

Zhu, J., Shen, Y., Zhao, D., and Zhou, B. (2020). "In-domain gan inversion for real image editing," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Glasgow), 592–608. doi: 10.1007/978-3-030-58520-4_35