Can Deepfakes be created on a whim?

Pulak Mehta
New York University
New York, New York, USA
pm3052@nyu.edu

Gauri Jagatap
New York University
New York, New York, USA
gauri.jagatap@nyu.edu

Kevin Gallagher
NOVA LINCS & Universidade NOVA de Lisboa
Lisbon, Portugal
k.gallagher@fct.unl.pt

Brian Timmerman
New York University
New York, New York, USA
brian.timmerman@nyu.edu

Progga Deb
New York University
New York, New York, USA
pd1353@nyu.edu

Siddharth Garg
New York University
New York, New York, USA
sg175@nyu.edu

Rachel Greenstadt
New York University
New York, New York, USA
greenstadt@nyu.edu

Brendan Dolan-Gavitt
New York University
New York, New York, USA
brendandg@nyu.edu

ABSTRACT

Recent advancements in machine learning and computer vision have led to the proliferation of Deepfakes. As technology democratizes over time, there is an increasing fear that novice users can create Deepfakes, to discredit others and undermine public discourse. In this paper, we conduct user studies to understand whether participants with advanced computer skills and varying level of computer science expertise can create Deepfakes of a person saying a target statement using limited media files. We conduct two studies; in the first study (n = 39) participants try creating a target Deepfake in a constrained time frame using *any* tool they desire. In the second study (n = 29) participants use *pre-specified* deep learning based tools to create the same Deepfake. We find that for the first study, 23.1% of the participants successfully created complete Deepfakes with audio and video, whereas for the second user study, 58.6% of the participants were successful in stitching target speech to the target video. We further use Deepfake detection software tools as well as human examiner-based analysis, to classify the successfully generated Deepfake outputs as fake, suspicious, or real. The software detector classified 80% of the Deepfakes as fake, whereas the human examiners classified 100% of the videos as fake. We conclude that creating Deepfakes is a simple enough task for a novice user given adequate tools and time; however, the resulting Deepfakes are not sufficiently real-looking and are unable to completely fool detection software as well as human examiners.

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04…$15.00
https://doi.org/10.1145/3543873.3587581

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**; **Social aspects of security and privacy**.

KEYWORDS

deepfakes, generative models, video synthesis

ACM Reference Format:
Pulak Mehta, Gauri Jagatap, Kevin Gallagher, Brian Timmerman, Progga Deb, Siddharth Garg, Rachel Greenstadt, and Brendan Dolan-Gavitt. 2023. Can Deepfakes be created on a whim?. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543873.3587581

1 INTRODUCTION

Deepfakes have gained attention both in technical literature as well as media. The concept of Deepfakes combines emerging techniques in *deep* learning to create *fake* multimedia content. It is a form of manipulated image, video, or audio data which can potentially be created with malicious intent. Deepfake content is typically created with the following two objectives i) to superimpose a target person's video onto a video of a source person and mimic its actions and ii) to make the target person emulate audio from the source person.

Pre-existing methods for generating artificial videos, use visual effects or computer graphics animation tools and require significant domain expertise to implement [28]. On the other hand, Deepfake creation tools have become easily accessible and require limited technical knowledge of the creation of fake multi-media content. The most popular platforms used are FaceSwap [1], ZAO [10] and Reface [8] among several others. DeepFakesLab, another Deepfake open-source software, provides a comprehensive package [34] for implementing several video manipulations such as face swaps, face aging, superimposing fake audio alongside lip syncing. FaceForensics [37, 38] provides an open-source Deepfake image dataset.

Deepfakes are typically created using generative neural networks like Variational Autoencoders (VAE) [23], Generative Adversarial Networks (GAN) [17] or flow-based generative models [22]. The

model is trained to explicitly learn the probability distribution of the image (or audio) training dataset. This allows the network to generate *unseen* images (or audio) which have characteristics similar to the images (or audio) in the training set. Due to the abundance of image and audio data available online these days, particularly for images of popular actors, politicians and entrepreneurs, it has become increasingly easier to train Deepfake generation models and produce realistic fake audio as well as videos.

1.1 Motivation

Deepfake's potential to deceive an average user gives it an edge over traditional image rendering. The Google search metrics [7] for the term "Deepfake" suggests a continuing upward trend in terms of interest among the general public. With the democratization of this technology, there is an imminent question: *can ordinary citizens generate Deepfakes with malicious intent in a very short time frame?* We further motivate this question as follows:

1.1.1 Accessibility. Platforms such as Faceswap [1] have given access to a common person with no domain skills to develop fake videos. Most applications [35, 47] allow anyone with a webcam and microphone to produce a video with the target person emulating the expressions, lip movements and speech of the source speaker. Along with the increasing processing power of GPUs and decreasing cost of cloud computing, these tools are becoming cost affordable [13].

1.1.2 Privacy and security. Over the past few years, social media and video-sharing platforms have been used with malicious intent to spread misinformation by impersonation of influential personalities. Deepfake videos targeting political figures include that of former United States (US) President Barack Obama [48] using an expletive to describe the then US President Trump. Similar examples have since surfaced online, without being marked as fake videos.

Several real videos have also been deemed as Deepfake, as a cover-up or deflection from political issues. When Gabon's president Ali Bongo [2] released a video addressing the public after months of having not been seen in public, Bongo's political opponents deemed the video as a Deepfake, leading to increased distrust among common public. There is a significant social impact of people viewing videos without knowing if they are fake.

Another area of risk is biometric authentication systems, like face authentication which has applications across Know-Your-Customer (KYC) compliance and fraud prevention. A recent study found that targetted deepfake attacks on face recognition tools achieved a success rate of 78% [46]

1.1.3 Curating trustworthy datasets. Generative models, which are the backbone of Deepfakes, have also been used to augment datasets for training machine learning models wherever there is insufficient training data [11]. Deepfake detection has a direct implication in curating trustworthy datasets.

1.2 Social impact

The ease and quality of Deepfake video generation point to several social challenges. Several papers have surveyed participants [15, 50] and evaluated their reactions to Deepfake news articles. They conclude that Deepfake news articles, particularly when combined with political microtargetting techniques reduced participants' trust in social media news outlets and also changed their perception of the target person involved. This can have serious implications ranging from outcomes of elections to inducing distrust in governments and the economy. There has been a notable amount of media coverage [33, 48] on the threat to democracy that Deepfakes may pose.

Due to the threat that Deepfakes may pose, several government administrations [12] have come up with laws that enforce regulations on Deepfakes, including the Deepfake Report Act under the 2021 National Defense Authorization Act (NDAA) of the US government. US state governments of Texas and California have also passed laws banning use of Deepfakes close to elections [12].

1.3 Limitations of current Deepfake platforms

It is important to note that the most easily available mobile applications such as FaceSwap [1], ZAO [10], and Reface, have limited use cases. They are targeted towards generating video fakes with fixed audio files added onto the target video. These audio files may be prerecorded clips from songs or movie dialogues and are only created for the purpose of entertainment. These videos generally conform to the policies of major social platforms and are not perceived as a risk to the society.

Through our discussion above, we note that most Deepfake content has high profile personalities. In this paper, we study a more challenging problem – that of creating a Deepfake with an ordinary person as the target. We then assess the risk it may pose, by using human and software detection tools.

In the next section, we outline the main questions addressed in this paper, as well as summarize key insights.

1.4 Our contributions

The existence of Deepfakes lead to several pertinent questions that we address in this paper. In particular, we probe the following

(1) Are Deepfakes easy to generate for a novice user[1] with varying skill levels?
(2) For the Deepfakes generated by our sample participants, are these Deepfakes generated realistic enough to fool
   (a) a set of human examiners
   (b) a Deepfake detection software tool?

We aim to answer these questions via our structured methodology that we highlight in detail in Sec. 4. Particularly, we design an objective where we pick both the source and target person. We assign two objectives to all categories of participants selected:

Objective 1.1. *(video) to transfer the expressions and lip movements of source speaker to that of target speaker*

Objective 1.2. *(audio) to transfer the speech of source speaker to that of target speaker.*

We conduct a two-fold study for the same; in the first part, we assign an open-ended objective and ask participants to use *any* Deepfake generating tool of their choice. In the second part, we ask the participants to use a *pre-defined* tools to create Deepfakes. We deem the task of creating a Deepfake as successful if the participants are able to synthesize a video with the target speaker uttering the

[1]By "novice" we refer to users who do not have direct expertise in creating Deepfakes.

target speech. Our study yields the following insights summarized below:

(1) Deepfake generation:
   (a) 23.1% of participants are able to complete both the Deepfakes Objectives 1.1 and 1.2 in the open-ended study.
   (b) 58.6% of participants are able to complete both the Deepfakes Objectives 1.1 and 1.2 successfully in the pre-defined tool study
(2) Deepfake detection: Deepfakes generated by participants are flagged as fake or suspicious
   (a) for 100% of Deepfake samples, when assessed by a set of human examiners.
   (b) for 80% of Deepfake samples, when analyzed by a software detection tool.

In the next section, we provide a comprehensive literature survey of topics relevant to our study.

2 PAPER ORGANIZATION

We have organized the paper as follows. In Sec. 3 (Related work) we discuss survey literature in related areas of Deepfake generation techniques (Sec. 3.1), Deepfake detection techniques (Sec. 3.2), User studies on Deepfake detection (Sec. 3.3) and User studies on Deepfake generation (Sec. 3.4). In Sec. 4 (Methodology) we outline the main structure of our two user studies; in particular we discuss in Sec. 4.1, details on participant background in Sec. 4.2 and participant recruitment in Sec. 4.3. We finally discuss the process of conducting the user study in Sec. 4.4.

In Sec. 5 we focus on the main findings of our user studies. We comment on the participant awareness about Deepfakes in Sec. 5.1. This is followed by main findings from our Open-ended Deepfake generation Sec. 5.2, Pre-defined Deepfake generation study Sec. 5.3. We finally present our Discussions and Conclusions in Sec. 6.

In Appendix A, we elaborate deep generative literature, details on participant compensation and demographics, ethical considerations, limitations of our approach and future direction. We also include some snapshot examples from deepfakes created in Appendix B.

3 RELATED WORK

We push the review on generative networks to Appendix A.

3.1 Targeted Deepfake generation

One of the first attempts at targeted Deepfake generation is [43], where the authors synthesize a high-quality lip-synced video of President Barack Obama given an input audio clip. The model was trained on nearly two million frames of Obama's weekly address footage using a recurrent neural network (RNN) that maps raw audio features to mouth shapes.

Face2Face [47] is another important benchmark and one of the first methods to successfully render fake facial expressions and audio, using a dense photometric consistency measure.

FaceSwap and DeepFaceLab are both open source software for Deepfake generation[1, 34]. Wav2lip [36] provides automatic lip-syncing given a target video and corresponding audio to be synthesized. It uses a pre-trained Discriminator, referred to as a "lip-sync expert" which allows their network to produce Deepfake video with high fidelity between audio and lip movements.

Another approach is first-order motion [42], where the Deepfake generation process is split into two parts; motion extraction via unsupervised key point detector, and generation which in-paints the target video along the key points extracted.

Some other recent papers include FSGAN, which does not require any training on the source or target subjects [32], RSGAN [30] which allows additional flexibility in explicit feature editing (like controlling the appearance of hair or eyes separately) and FaceShifter [25] which improves the fidelity of generated Deepfake video under facial occlusions.

It is worth noting that the code for implementing most of these Deepfake generation techniques [30, 32, 36] is open-source and requires the user to just have access to a computer with basic hardware requirements. Platforms like Google Colaboratory also provide free access to limited GPU computing.

3.2 Deepfake detection techniques

The Deep Fake Detection Challenge (DFDC) [16] consists of $\sim$ 124, 000 videos and featured eight facial modification algorithms. The winning detection technique from the challenge by Seferbekov [40] consists of breaking the video down frame by frame and using a multi-task cascaded convectional network (MTCNN) for face detection [51] and extracting a crop that excludes the background. This is followed by running an EfficientNet [45] classifier to detect the face crop as real or fake.

Deepware [5] is an API that directly takes a video (fake or real) from the user and tests against four different machine detection algorithms. The four detectors considered are: i) avatarify [3] which is a face animation app, ii) deepware [5], iii) the MTCNN and EfficientNet methodology by Seferbekov [40], and iv) an ensemble method that combines deepware detector and Seferbekov's detector. Other techniques for detection that are not covered by Deepware API include but are not limited to: Mesonet, which utilizes mesoscopic image information to build a neural classifier [9] and Face-Cutout [14] which is a data-augmentation procedure which improves upon Deepfake detection accuracy of prior art.

We refer the reader to [29, 31, 49] for a more comprehensive overview of both Deepfake generation and detection techniques.

3.3 User studies on Deepfake detection

In academia, there is no clear consensus on whether human participants can successfully distinguish between real and fake videos.

[44] provides a comprehensive study on human participants equipped with the task to assess if provided Deepfakes were real or fake. Their survey suggests that majority of the people are unable to classify a Deepfake image to be fake.

The study in [24] suggests that participants had a systematic bias toward guessing that videos are authentic, fake or real. They conjecture that people apply an overly optimistic seeing-is-believing heuristic, which points to a larger risk of the common public being influenced by Deepfakes.

The study in [18] suggests that Deepfake detection software and human examiners have similar classification accuracies on the same dataset, however, make different mistakes when misclassifying a media. They propose that a combined classification procedure that

uses both machine detection and human visual examination is the best strategy when it comes to Deepfake detection.

We refer the reader to [26] which characterizes the shortcomings of the current state of Deepfake detection techniques.

3.4 User studies on Deepfake generation

To our knowledge, there is no other paper that discusses the capability of novice users to make Deepfakes. We now present the main setup of our user study in the next section.

4 METHODOLOGY

We start with outlining the various considerations that went into designing and conducting the two-part user study on Deepfake generation. See Figure 1 for a visual representation of the study.

4.1 User study

We devise a user study to understand Deepfake generation by requiring a set of pre-selected participants to create an audio-visual Deepfake of a target person speaking a target statement. This is the most general case of Deepfake generation and can have major social-economical ramifications. In our study, the target person was Kevin who is a project member and the target statement was:

"Hello everyone, my name is Kevin and this is a Deepfake video".

Standards techniques for Deepfake generation techniques rely on transferring the lip and facial movements from one video clip onto another video. We provide the following media files to the novice user to create a Deepfake: (i) video of the target person and (ii) a video clip of source speaker speaking the statement that the target needs to emulate.

The task is assessed as successful if the output of the fake audio-visual used a Deepfake generation tool to create the audio and video components along with stitching them together.

The participants were provided with three videos, each being $\sim$ 1 hour long, of the target person (target video style) speaking in their voice (target audio style). Additionally, they were provided with a video clip of another research participant (source video) speaking a $\sim$6 second clip of the target statement in their voice (source audio).

We conducted four pilot studies to iterate and determine the best structure for the user study. These pilots consisted of two participants from the research team who have experience with deep machine learning techniques and two participants with limited machine learning knowledge but from a computer science background. The pilot participants were asked to create Deepfakes on a whim with the same media files. Three of the participants were able to complete the task within 1.5 hours with no guidance, while the last participant finished within 2 hours with limited guidance on debugging the Deepfake creation tool. The biggest challenge faced by the pilot participants was identifying the right tool, followed by debugging while using these tools.

Based on this, considerations for designing the Deepfake generation user study are as follows:

4.1.1 Time frame. In the pilot study three out of four participants needed less than 1.5 hours implying that generating Deepfakes in a time-bound manner is possible. Moreover, we also want to assess how practical it is to make Deepfakes *on a whim*. Studies [39] suggest that certain Deepfakes can be created in less than 6 minutes. Based on these considerations, we assign a time limit of 2 hours for both parts of the study.

4.1.2 Ease of accessibility of Deepfake generation tools. We investigate if the identification and accessibility of the Deepfake generation tool is a major challenge in the creation of Deepfake content. We design a two part study, one where tools for creating Deepfakes are not provided, and one where they are.

4.1.3 Ease of implementation of Deepfake tools. We construct a survey to assess the difficulties experienced while implementing Deepfake algorithms.

These factors lead us to create the two parts of the study.

The first part of the study required the participants to create Deepfake using *any tool*. We refer to this part of the study as *Open-ended Deepfake generation study* for the remainder of this paper.

For this part of the study, participants are free to use search engines, mobile applications, online software tool, or reuse any code base. They are allowed to use their personal computers or available cloud servers to generate the desired fake video. The study is designed to be representative of Deepfake creation in the real world. To set a control parameter for the study, we allot all participants 2 hours for the first part of the study.

In the second study, we re-invite the same participants for creating the same target Deepfake using *pre-specified tools* for fake audio and video generation in a time frame of 2 hours. We subsequently refer to this part of the study as *Pre-defined Deepfake Generation study*. The participants are also provided with publicly available tutorials for using these tools. More details about the tools are provided Sec. 5.3.1.

Using this two-fold approach, we investigate whether Deepfake generation tool identification is a major challenge in the generation of Deepfakes. The second part of our study enables us to examine if given the correct set of tools, can novice users generate Deepfakes.

Lastly, we run both software-based and human-examination-based detection on the fakes generated to investigate how many of the videos generated are actually *realistic*.

4.2 Participant background

Generation of audio or video Deepfakes from scratch requires skills such as software, machine learning and basic media editing like cropping, merging media etc. Additionally, video Deepfakes can be enhanced using Visual Effects (VFX) techniques [27]. Hence, for our study, we select four categories of participants that represent the broad skill sets required to create fakes. These categories are-

(1) Basic computer skills - The only coding coursework completed by the participant is introduction level courses on computer science, during their academic coursework.
(2) Intermediate computer skills - Participants have studied multiple computer science or coding courses in their coursework, but not machine learning.
(3) Advance computer skills - Participants have taken at least one machine course during their coursework.
(4) Digital Media skills - They have completed at least one coursework related to Visual Effects.

All participants surveyed have a basic understanding of a computer either through an introduction to computer science course or

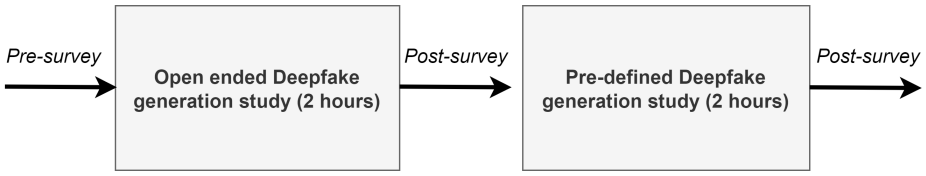

Figure 1: Overview of the two-part user study. The pre-survey captured participants' awareness of Deepfakes.

a Visual Effects class and possess the ability to perform basic media file edits, such as cropping, merging etc. or can quickly grasp the concepts for the same during the experiment session. We received a few participants with overlapping categories of basic computer skills and digital media skills. They were placed under digital media skills for the study.

4.3 Recruitment

Recruitment emails were sent to the first author's university mailing list stating that this study is regarding Deepfakes. The email did not contain any information about the actual tasks that the participants would be required to do for the study. To sign-up for the study, participants provided their details such age, degree program enrolled, relevant coursework, gender, and ethnicity along with their consent for the voluntary study. We attempted to select a diverse set of participants across each of the participant categories. We expected dropouts as the study had multiple parts.Hence we invited 52 participants (13×4 categories) for the study. Out of the 52 participants, only 39 turned out for the open-ended tool study, and 29 for the pre-defined tool study. The breakdown of participants that were finalized is described in Appendix A.3.

4.4 Study procedure

The sessions were conducted virtually using Zoom in compliance with the health safety advisor by the local authorities. At the start of a session, participants switched on their webcams for the duration of the sessions and were allotted a unique identification code which remained the same for both of the studies. They were required to change their display name to the identification code.

For the first session, a pre-survey was conducted to understand their knowledge of Deepfakes. After the pre-survey, participants were provided with the task for that session. All participants attempted the Deepfake creation task in the allotted time of the corresponding study. During each study, we provided assistance through a stop point survey (see Figure 1). If participants faced any challenge for more than 10 min and required guidance for the next step, they were instructed to fill out a brief stop-point survey highlighting their problem. A member of the research team was then assigned to guide them in solving the problem in a breakout room. The research team only provided direction and did not debug the code or assist in solving the problem completely. After the allocated time, the participants filled out a brief post-survey alongside submitting their outputs.

In Table 1 and Table 2 we break down the survey rubric for the pre-study survey and post-study survey respectively.

We describe details about participant compensation and ethical considerations in Appendix A.

Table 1: Breakdown of survey rubric prior to user study.

Rubric	Questions
Awareness	Heard about Deepfakes prior? Sources of information?
Knowledge	How knowledgeable about Deepfakes?
Expertise	Have they created Deepfakes before? How optimistic about finishing objective?

Table 2: Breakdown of survey rubric post user study.

Rubric	Questions
Techniques	Broad strategy to solve objective? Most challenging step? Time spent on each step? Software tools used?
Confidence	How optimistic about creating Deepfakes? Did they lose motivation during study?
Task completion	Fake video created? Fake audio created? Fake audio and video combined? Deepfake video realistic?

5 FINDINGS

5.1 Awareness around Deepfakes

We collected insights about Deepfakes awareness from the pre-survey (Table 1) at the start of the open-ended Deepfake generation study. Majority of the participants were aware of Deepfakes (∼84.6%) before the recruitment email, see Figure 2. These participants reported having heard about Deepfakes across various platforms. The most popular platforms were social media sites (n=23), closely followed by online video sites (n=22), news outlets (n=19), friends and family (n=17), academic work (n=14), and messaging apps (n=4) in decreasing order. The most common social media website where participants encountered Deepfakes was Instagram and TikTok, while the most common online video site was Youtube.

Additionally, a small number of participants (∼20.5%) reported having seen Deepfakes on the internet without a Disclaimer stating that they were Deepfakes (see Figure 3). This can be a cause of concern when Deepfakes, when misinformation targeted fake, is available online without disclaimers.

At the start of the study, only 3 out of the 39 participants (∼7.7%) had previously attempted to create Deepfakes using free mobile apps with the objective of creating memes/ humorous content.

5.2 Open-ended Deepfake generation study

5.2.1 Overview. In this session, participants were required to generate the target Deepfake in 2 hours without being recommended

Table 3: Breakdown of technical expertise for participants in Open-ended Deepfake generation study. We ensure that number of participants in each category is roughly equal.

Category	Participants
Basic computer skills	10
Intermediate computer skills	10
Advance computer skills	11
Visual Media Skills	8
Overall	39

Table 4: (Stop point survey from open ended study) Breakdown of the key challenges encountered to generate a Deepfake . Most of the participants encountered challenge in debugging the code on Google Colab.

Challenge type	Percentage
Debug - Google Colab	**35.8**
General query	17.9
Debug - Local computer	14.3
Tool identification	14.3
Tool taking time	10.7
Media editing	7.1

any specific tool or tutorial. Participants were provided all the media files (three of the target person and one of the source persons speaking the target statement). Out of the 41 participants who started the study, two of them dropped mid-way. One of them was worried that the study would require them to download software and they may accidentally download malicious software. A second participant dropped out due to unknown reasons. Overall 39 participants completed the session (a breakdown of the technical expertise of these participants is presented in Table 3).

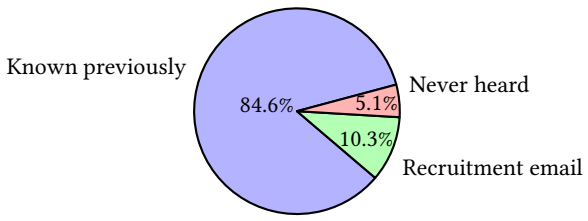

Figure 2: Deepfake awareness among participants.

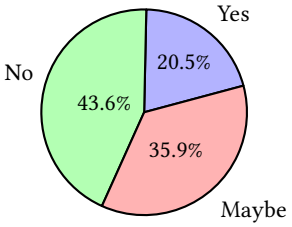

Figure 3: Did the participants see Deepfakes *without* any disclaimer?

Table 5: (Output from open-ended study) Based on human review, we find that 23.1% of participants are able to create a Deepfake. These numbers do not represent if the Deepfake created can be detected as a fake by the human examiners or software detector.

Deepfake component created	Percentage
Deepfake Video without audio	15.4%
Deepfake Audio without video	10.2%
Deepfake audio and video	23.1%
Neither audio nor video	51.3%

5.2.2 Stop point survey. We received 28 requests for support with Deepfake generation using the stop-point survey. 50% of the issues were related to generating video fakes, 35.7% were related to generating audio fakes and the remaining were general issues like clarification of the study's objective. Hence, even with limited or no knowledge of making fakes, most people were able to make headway in creating them. However, not all participants identified the correct set of tools to implement the task. Most of the participants faced challenges in debugging the identified tool. These predominately arose in debugging tools running on Google Colaboratory, a platform that allows users to run Python code. We highlight these findings in Table 4.

The motivation of participants also varied during the session. 60% of the participants reported that they had sustained motivation throughout the study. The loss in the motivation of the other 40% was attributed to factors such as long debug time, lack of coding or machine learning knowledge etc. Interestingly, some participants reported that they became more motivated as time progressed. In terms of time split while creating Deepfakes, ~56.3% of the time by participants was spend on video fakes, ~26.4% on audio fakes and ~20.3% on other activities like tool identification, media editing etc.

5.2.3 Deepfakes generated. Within a two-hour period, a sizeable number of the participants were able to generate Deepfakes. The outcomes are highlighted in Table 5.

Our criteria of whether a given participant created a Deepfake successfully, is whether they used an existing Deepfake tool to create the audio and video. This meaningful completion rate of 23.1% can be attributed to Deepfake tools being made publicly available on Google Colab with detailed instructions by the authors. We add details about techniques used in Deepfake generation in the open-ended study as follows in Tables 6 and 7.

After generation, the next step is to analyze if the Deepfake quality of the video is good enough to mislead humans or computers.

5.2.4 Deepfakes detection. Most of the work around Deepfake detection revolves around detecting video, and not the audio component. We process all the videos through a Deepfake detector to identify whether it can bypass detectors. We leverage the Deepfake detection API provided by Deepware[5] to quantify the quality of the fake. Their tool uses four major algorithms to determine if video is fake or not - Seferbekov[40], Avatarify, Deepware and an ensemble method that leverages the power of the previous three methods. Seferbekov's method [40] is the winner of the popular Deepfake Detection Challenge (DFDC)[16], whereas Avatarify which is aimed at detecting fakes generated using the Avatarify tool. The output

Table 6: (Open-ended study). Breakdown of techniques used by 15 people who created fake video successfully. These numbers do not represent if the Deepfake created can be detected as a fake by the human examiners or software detector.

Method	Details	Num. participants
First Order Motion[42]	Academic open source	5
Speakr	Mobile app	2
Reface	Mobile app	2
DeepFakeLab [34]	Open source	1
Wav2lip [36]	Academic open source	1
Adobe After effects	Image enhancement software	1
DaVinci	Image enhancement software	1
Snapchat	Mobile app	1
Avatarify [3]	Web tool	1
Total		15

Table 7: (Open-ended study). Breakdown of techniques used by 13 people who created fake audio successfully. These numbers do not represent if the Deepfake created can be detected as a fake by the human examiners or software detector.

Method	Details	Num. participants
SV2TTS [21]	Text-to speech open source	9
DeScript	Web tool	1
Tacotron 2 [41]	Text-to speech open source	1
Mozilla TTS	Text-to speech open source	1
Total		13

Table 8: Of the 48.7% videos that were partially or completely generated as Deepfakes, we use the Deepware [5] software tool to flag the videos as real or fake. We find that most of the Deepfake videos created by participants are flagged as fake using Deepware.

Video type	Percentage
Detected fake	**53.3%**
Detected suspicious	26.7%
Not detected	20%

of the analysis for the Deepfakes created in the open-ended study is provided in Table 8.

The Deepware API tool processes video and each of the four algorithms reports a score from 0-100, where 100 is fake. The suspicious videos had an average score algorithmic score between 52-72 by each of the four algorithms. The videos that were flagged as "not fake" video had the best algorithm report a score of 48 and 49. There are many factors that we observed that determined the fakeness by this detector. This included the number of faces in a video, face movement, face direction etc. An interesting observation was that one student from the digital media background used DaVinci resolve, which is a video editing application, to manipulate the lip movements. Such a process is extremely time-consuming to create a realistic fake, however, the best out of the four algorithms provided it with a score of 7. Hence, future detectors should be trained to identify fakes that can be generated using video editing software like DaVinci Resolve.

We further conducted a human-based analysis of the videos generated. It consisted of five examiners of the following background -
1) 20 year old, male, social science background
2) 23 year old, female, physical science background
3) 24 year old, male, physical science background
4) 27 year old, female, journalism background
5) 29 year old, female, engineering background
The human analyzers gave us their consent; however were not financially compensated. They reviewed the outputs from the open-ended study on a laptop and all values reported by this group hereafter have been averaged out. Note that the users were informed before-hand that the videos they view may be real or fake.

We instruct the examiners as follows:

(1) detect if the video with muted audio is fake or real
  (a) if detected as fake, comment on which parts of the video (including facial features) has distortions.
(2) detect if the audio without the video sounds
  (a) like human
  (b) vaguely like target
  (c) similar to target
(3) detect if audio and video played together looked
  (a) lip-synced
  (b) realistic; i.e. no video distortions; sounds similar to target.

The reviewers reported only $\sim$ 25.6% of the videos had the face looking like target person. The other videos included superimposing select features of source like eyes, mouth onto Kevin's hair and beard, massive distortion etc. Such details may bypass a programmatic detector but can be easily identified by the human eye.

Analysis of human examination of the Deepfakes generated in the open-ended study is in Table 9. Similar to [44], we dissect the distortions in different zones of the video, such as eyes, lips, etc. Only found two videos were flagged as real with muted audio.

Table 9: Human expert based visual inspection of properties of Deepfake video that appear unreal and may cause the Deepfake video to be flagged as fake. In most Deepfake videos we find that the lips/mouth movement to be the most distorted (highlighted).

Video Area	Videos not Distorted
Eyes	82.05 %
Facial hair	76.92 %
Lips/Mouth	**69.23** %
Nose	74.36 %
Cheeks	71.79 %
Forehead	74.36 %
Hair	71.79 %
Background/other	82.05 %

We also conducted an audio analysis of these fakes by human reviewers. The output of most fakes sounded like human. Our analysis of the audio clips with no video is highlighted in Table 10.

Upon watching the audio and video together we found only 15% of the videos to be vaguely lip-synced. However, none of the fakes were realistic enough to demonstrate that the target person was speaking the target statement.

Table 10: Human based auditory inspection of Deepfake audio quality. Most Deepfake videos have audio that sounds like a human, but not necessarily like target person.

Audio Details	Percentage
Sounds like human	65.24%
Sounds vaguely like target	34.76 %
Sounds similar to target	0.0 %

Table 11: Breakdown of technical expertise for participants in the pre-defined Deepfake generation study.

Category	Participants
Basic computer skills	7
Intermediate computer skills	9
Advance computer skills	7
Visual Media Skills	6
Overall	**29**

5.3 Pre-defined Deepfake generation study

5.3.1 Overview. In this session participants were asked to generate the target Deepfake in 2 hours using a specific tool. The tools that we provided for creating fakes were the Google Colaboratory version for the Real-Time Voice Cloning [21] and Wav2lip [36] for audio and video respectively, which are state-of-art open source repositories. We provided access to condensed tutorials for the audio [4] and video [6] based on reference papers [21, 36]. Additionally, the Google Colaboratory version of these tools ensured that all participants had access to similar hardware.

All participants were given the same media files as the prior session. Due to dropouts, only 29 completed this study (breakdown of their technical expertise is presented in the Table 11).

5.3.2 Stop point survey. We received 15 requests for support during the study which we captured through the stop-point survey. Since the tools and tutorials were provided most issues were either related to debugging Google Colaboratory (10), general queries (3), or editing media (2). We observed that Safari browsers, even with the latest version, were facing problems but as soon as they switched to Google chrome their problem got resolved.

5.3.3 Deepfake generated. To generate the Deepfake, participants first created the audio fake using the Real-Time Voice Cloning [21]. Then the participants crop a video file of length equal to the length of the audio fake. Finally, the audio fake is lip-synced onto the video clip using Wav2Lip. Participants used the above methodology and generated the Deepfakes highlighted in Table 12.

5.3.4 Deepfakes detection. Since all the fakes were generated using the same tools, they have similar fingerprints introduced by the Deepfake generator. We ran the same detection techniques on all the fakes and the summary is as follows -

(1) Detection algorithms provided by Deepware's API detected all videos as suspicious. Seferbekov algorithm reported a score of 95-99 (100 is fake), while Avatarify and Deepware algorithms reported score between 17-32.
(2) Human analyzers unanimously agreed that the videos with muted audio appeared realistic without any distortions

Table 12: (Output from pre-defined study) Based on human review, we find that 58.6% of participants are able to create a Deepfake. Note that these numbers do not represent if the Deepfake generated can be detected as a fake by the human examiners or software detector.

Deepfake component created	Percentage
Deepfake audio and video	58.6%
Deepfake audio only	31.0%
Neither audio nor video	10.4%

(3) Human analyzers unanimously agreed that the fake audio without video sounded vaguely like Kevin
(4) Human analyzers unanimously agreed that all videos were slightly out of sync in parts when examined very carefully.

6 DISCUSSION AND CONCLUSIONS

In this paper, we conducted a systematic study on the accessibility of Deepfake creation technology.

Deepfake awareness. Majority of our study participants (84.6%) have heard and encountered Deepfakes on social media websites or video sharing platforms. This is a positive step as awareness is the first step in fighting the challenges posed by Deepfakes.

Deepfake generation is not difficult for a novice. A substantial number of novice users 58.6% were able to generate Deepfakes successfully in a limited time frame, given appropriate tools. Even without guidance on which tools to be used, 23.1% of the novice users were able to generate fakes. This ease in generation can be attributed to Deepfake generation code being open-sourced on Google Colaboratory and instructions being readily available.

Quality of Deepfakes created. We note that of the Deepfake videos created by participants successfully from the open-ended study, roughly 53.3% were flagged as fake and another 26.7% are flagged as suspicious by our chosen software detection tool. This implies that the Deepfake generated by novice users are not realistic enough to fool a software detector. We also use the input of a set of five human reviewers. Even in this case, we find that 100% videos are flagged as fake based on a combined audio-visual inspection. However, from the pre-determined tool study, all the videos were detected as suspicious by the same detection tool. Even though 100% of the videos were flagged as fake by human reviewers, none of them demonstrated any visual distortion in the video component unlike the outputs from the open ended study. Additionally, all audio fakes sounded vaguely like the target person, unlike 65.3% which sounded like human but not vaguely like the target person in the open-ended study. Hence, we achieved better quality fakes in the pre-determined tool study, although we highlight that existing available deepfake creation systems still fall shot at fidelity.

Key challenges in generating fakes. Our study demonstrated that the key challenges are the identification of the right tool followed by debugging or using the tool. This can be demonstrated by the jump in the fakes created between the two studies. Additionally, there are many tutorials publicly available to help guide this journey.

We discuss limitations and future directions for our work in Appendix A.

REFERENCES

[1] 2017. Faceswap: Deepfakes Software For All. (2017). https://github.com/deepfakes/faceswap.

[2] 2019. Deeptrace: The state of Deepfakes. Landscape, Threats and Impact. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.

[3] 2020. Avatarify: AI face animator. https://avatarify.ai/.

[4] 2021. Audio Deepfake Tutorial. (2021). https://web.archive.org/web/20210422160657/https://www.youtube.com/watch?v=1WN8Jhfd4uM.

[5] 2021. Deepware: Scan and detect Deepfake videos. (2021). https://scanner.deepware.ai/.

[6] 2021. Video Deepfake Tutorial. (2021). https://www.youtube.com/watch?v=Ic0TBhfuOrA.

[7] 2023. Google search metrics for Deepfakes. https://trends.google.com/trends/explore?date=today%205-y&q=Deepfake.

[8] 2023. Reface app. (2023). https://hey.reface.ai/.

[9] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 1–7.

[10] Alexandros Antoniou. 2019. Zao's deepfake face-swapping app shows uploading your photos is riskier than ever. *The Conversation* (2019).

[11] Antreas Antoniou, Amos Storkey, and Harrison Edwards. 2018. Augmenting image classifiers using data augmentation generative adversarial networks. In *International Conference on Artificial Neural Networks*. Springer, 594–603.

[12] Scott Briscoe. 2021. U.S. Laws Address Deepfakes. (2021). https://www.asisonline.org/security-management-magazine/latest-news/today-in-security/2021/january/U-S-Laws-Address-Deepfakes/.

[13] Aengus Collins. 2019. *Forged authenticity: governing deepfake risks*. Technical Report. EPFL International Risk Governance Center (IRGC).

[14] Sowmen Das, Selim Seferbekov, Arup Datta, Md Islam, Md Amin, et al. 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*. 3776–3785.

[15] Tom Dobber, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes de Vreese. 2021. Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics* 26, 1 (2021), 69–91.

[16] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397* (2020). https://ai.facebook.com/datasets/dfdc/.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[18] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences* 119, 1 (2022).

[19] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. 2021. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

[21] Corentin Jemine et al. 2019. Master thesis: Real-time voice cloning. (2019).

[22] Durk P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems* 31 (2018). https://openai.com/blog/glow/.

[23] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[24] Nils C Köbis, Barbora Doležalová, and Ivan Soraperra. 2021. Fooled twice: People cannot detect deepfakes but think they can. *Iscience* 24, 11 (2021), 103364.

[25] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2019. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457* (2019).

[26] Neal Mangaokar and Atul Prakash. 2021. Dispelling Misconceptions and Characterizing the Failings of Deepfake Detection. *IEEE Security & Privacy* 01 (2021), 2–8.

[27] Neal Mangaokar and Atul Prakash. 2021. Dispelling Misconceptions and Characterizing the Failings of Deepfake Detection. *IEEE Security Privacy* (2021), 2–8. https://doi.org/10.1109/MSEC.2021.3099782

[28] Terrence Masson. 1999. CG 101: a computer graphics industry reference. (1999).

[29] Yisroel Mirsky and Wenke Lee. 2021. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)* 54, 1 (2021), 1–41.

[30] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. 2018. Rsgan: face swapping and editing using face and hair representation in latent spaces. *arXiv preprint arXiv:1804.03447* (2018).

[31] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Cuong M Nguyen, Dung Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. 2019. Deep learning for deepfakes creation and detection: A survey. *arXiv preprint arXiv:1909.11573* (2019).

[32] Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7184–7193.

[33] Simon Parkin. 2019. The rise of the deepfake and the threat to democracy. https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy.

[34] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. 2020. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535* (2020).

[35] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2018. Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. In *International Conference on Learning Representations*.

[36] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 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*. 484–492.

[37] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2018. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179* (2018).

[38] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1–11.

[39] Hrutuja Satpute, Samruddhi Raut, Pragati Sapke, and Mrudul Dixit. 2020. Deepfake Creation Using Generative Adversarial Network. (2020).

[40] Selim Seferbekov. 2020. Deep Fake Detection Challenge (DFDC) Solution. (2020). https://github.com/selimsef/dfdc_deepfake_challenge.

[41] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4779–4783.

[42] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First order motion model for image animation. *Advances in Neural Information Processing Systems* 32 (2019).

[43] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–13.

[44] Rashid Tahir, Brishna Batool, Hira Jamshed, Mahnoor Jameel, Mubashir Anwar, Faizan Ahmed, Muhammad Adeel Zaffar, and Muhammad Fareed Zaffar. 2021. Seeing is Believing: Exploring Perceptual Differences in DeepFake Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[45] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[46] Shahroz Tariq, Sowon Jeon, and Simon S. Woo. 2022. Am I a Real or Fake Celebrity? Evaluating Face Recognition and Verification APIs under Deepfake Impersonation Attack. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 512–523. https://doi.org/10.1145/3485447.3512212

[47] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2387–2395.

[48] Rob Toews. 2020. Deepfakes Are Going To Wreak Havoc On Society. We Are Not Prepared. https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/?sh=4b68fb7a7494.

[49] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion* 64 (2020), 131–148.

[50] Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society* 6, 1 (2020), 2056305120903408.

[51] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503.

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.

A APPENDIX

A.1 Deep generative networks literature

The most popular generative model used in literature is Generative Adversarial Networks (GANs) [17]. The network consists of two parts - the first one being a Generator and the second one being a Discriminator. The task of the Generator is to take a low-dimensional noise vector and map it to a realistic looking "unseen" image that matches the probability distribution of the training set. The task of the Discriminator is to successfully distinguish between "fake" images generated by the Generator and real images from the training dataset. The task of the Generative model is to successfully fool the Discriminator into classifying its outputs as "real" images. In the process, the Generator *learns* to produce more and more realistic-looking images. Furthermore, the fake images can be *conditioned* to look like a reference input image , such as in Pix2Pix [20], which is a generative model that enables image to image transformation.

Pix2Pix and CycleGAN [52] are popularly used for *style transfer* where the task is to transfer the style characteristics of the source image to the target image. Conditional GANs such as the Glow model [22] allow a user to condition the output of the generator on user-defined parameters such as hair color, eye shape, and age among others. Glow also allows a user to morph seamlessly between two different face images.

Note that deep generative networks have been used successfully for various applications such as producing realistic animation, image and video editing such as inpainting, denoising, 3-D model generation, lossless multi-media compression [19] apart from Deepfake generation.

A.2 Participant compensation

Each candidate was compensated for their time and effort. For the open-ended Deepfake generation study, they were awarded 35 USD for their time and a 10 USD bonus if they are able to create a complete audio-video fake. For the pre-defined Deepfake generation study they were awarded 45 USD for their time and 10 USD bonus if they are able to complete audio-video fake. This compensation was structured to reward their time, increase motivation during the session and reduce dropouts. All the compensation was paid out in the form of gift cards from a leading e-commerce company.

A.3 Participant demographics

We chart out the demographics of the participants who actually participated in our study in terms of (a) gender (female, male, non-binary), (b) age range (18–20,21–23,24–26,27–29,30–39,40–49) and (c) race (White, Black or African American, Asian-Indian, Asian-Chinese, Asian-Korean, Asian-Other, and Hispanic). We plot numbers corresponding to the *open-ended* study in blue and those corresponding to the *pre-defined* objective in red and represent our data in Figure 4. We note that a total of $n = 39$ participants were present for the first part of the study and $n = 29$ participants were present for the second part. Breakdown of technical expertise for participants in both studies is also given in Table 3 and 11 respectively.

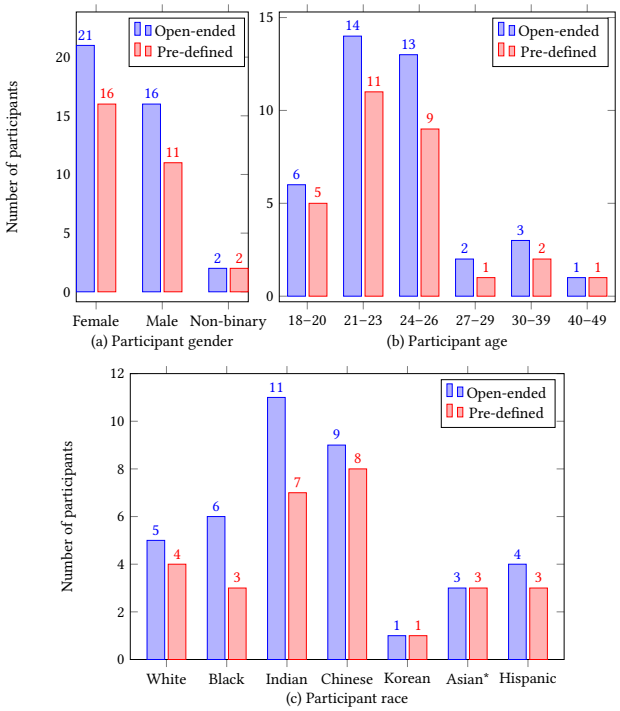

Figure 4: Demographic breakdown of participants, in terms of (a) gender, (b) age, (c) race (*Asians not counted under Indian, Chinese or Korean). We plot numbers corresponding to the first study (with open-ended objective, n=39 participants) in blue and those corresponding to the second study (pre-defined objective, n = 29 participants) in red.

A.4 Ethical consideration

This study was approved by the Institutional Review Board (IRB) of the first author's university. All participants were required to be over the age of 18 and provided consent for the study. We provided necessary caution to all participants at the start and end of each of the studies that the objective of this research study was to understand the current state of Deepfakes from an academic perspective only. Participants were strongly suggested to use any knowledge gained about Deepfakes from this study judiciously. Participants were also directed to delete all the Deepfake-related files from their local computer or cloud after the session ended.

A.5 Limitations

A remote study helped us in understanding Deepfake generations while simultaneously ensuring health safety; however, such study design comes with several limitations. Our participants sample represented young adults who are pursuing both undergraduate and graduate degrees. We relied on participants self reported their coursework to recruit and assign groups. The media files provided of the target person had all the footage with the face looking straight into the camera. Such front-facing video is optimal for creating fakes and may not be always available when creating fakes in the real world. The participants were also limited by the capabilities of their Operating system, hardware and internet speed. Many Deepfake

tools run only on Microsoft Windows, require GPU such as NVIDIA CUDA and require downloading files that could be as large as a few gigabytes. All these factors can limit the Deepfake generated by participants. The time-bound nature of the study is another factor that can inhibit the results. Many participants reported in the post survey that they would be able to create the fakes if more time was provided. Lastly, self-reporting surveys have biases such as social desirability biases.

A.6 Future directions

Based on our study and conclusions, we propose some key steps to further the findings of our analysis.

Revising the parameters of our study: For future, we propose to conduct this study with a larger sample set that is representative of a diverse population. One of the limitations of our study was also the time constraint under which participants had to create the Deepfake. Instead of two hours, one may increase the time frame of the study and infer the results.

Social impact and introducing restrictions: Since we survey the impact of increased accessibility of Deepfake software, another aspect of the study may involve inspecting the implications of sharing Deepfakes online through social media websites. We can also investigate guidelines and restrictions on sharing such content online.

Easier access: The task of creating Deepfakes may be outsourced to external agencies or artists, instead of novice users. We can evaluate the quality of Deepfakes generated by experts under no time constraints and compare them against the data collected for the users in this study.

B SUPPLEMENTARY RESULTS

The results obtained from both studies exhibited diverse variations. The output was contingent upon the techniques employed for image/video cropping, and, the Deepfake generation tool used. Thus, demonstrated significant disparities. The input media files had the target person facing the camera, which considerably facilitated the generation of Deepfakes. Participants adopted varied cropping strategies, with some confining themselves to the facial region as seen in fig. 5 and others encompassing the upper body as seen in fig 6. Our anecdotal findings suggested that automated deepfake detectors exhibited greater proficiency in detecting videos containing only the face, rather than those encompassing the upper body. Moreover, we noted that the output generated from the first study displayed more distortions (see fig. 7) in comparison to that of the second study (see fig. 8). This discrepancy could be attributed to the restricted toolkit available in the latter study.

B.1 Acknowledgments

This work was supported in part by NSF 2016061.

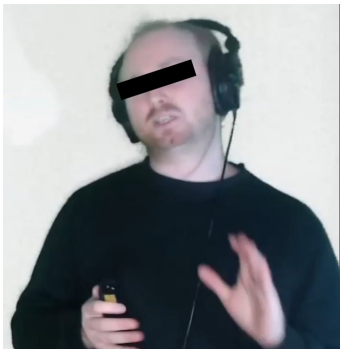

Figure 5: A realistic Deepfake produced from study-1

Figure 6: A realistic Deepfake produced from study-2

Figure 7: A distorted Deepfake from study-1

Figure 8: A distorted Deepfake from study-2