LUISS

Department of Business and Management

Course of Customer Intelligence & Big Data

Deepfakes: Analysis on the Role of Disclosure Placement in Consumers' Attitude Towards Synthetic Advertisement

Prof. Giuseppe Francesco Italiano

SUPERVISOR

Prof. Daniele D'Ambrosio

CO-SUPERVISOR

Andrea Parialò - 725281

CANDIDATE

Academic Year 2020/2021

Table of Contents

1. **Introduction 6**
2. **History of Deepfake 6**
3. **Different types of deepfakes 12**
    1. Facial Re-enactment 12
        1. Face2Face 12
        2. Neural Voice Puppetry 13
    2. Face Swapping Deepfakes 13
    3. Full Body Deepfakes 14
    4. Audio deepfakes 14
4. **Concerns, threats, risks: 15**
    5. Fake News 15
    6. Democracy 17
    7. Pornfakes and cyberbullism 19
    8. Scams and impersonifications 19
    9. The Liar's Dividend 20
    10. Visual Dominance 21
5. **Solutions and countermeasures to deepfakes: 22**
6. **A technical overview and discussion. 22**
    11. Detection algorithms 22
    12. Blockchain architecture 25
        3. Smart Contracts 25
        4. Hyperledger Fabric 25
        5. Why the blockchain? 26
7. **Opportunities: when deepfakes are a godsend 27**
    13. Edutainment 27
    14. Accessibility 28
    15. Therapy and mental health 28
    16. Removing Language Barriers 28
    17. Freedom of speech 29
    18. Film Industry 30
    19. Gaming industry 31
8. **A big deal for marketing 32**
    20. Virtual Influencers 33
    21. The Uncanny Valley Effect 34
    22. Bigger reach 34

23. Hyper-personalized ads 35
24. Tourism 36
25. The Metaverse 37
9. **Risks for Marketing** **37**
26. Data use and privacy 37
27. Perceived reality 38
28. Ad falsity perception and verisimilitude 38
29. Government policies 40
30. Consumers' reaction to low budget ads 40
31. General perception 40
10. **Research Question** **41**
11. **Research Methodology and Design** **42**
32. Choosing and editing the media 42
33. Identification code 46
34. Survey building 47
6. Pre-test 47
7. Main Test 50
12. **Data Analysis** **53**
35. World map 54
36. Pre-Test 54
8. Pre-test demographics 54
9. Pre-test results 55
37. Main Test 55
10. Demographics overview 55
11. PCA (Principal Component Analysis) – Loading Plot 56
12. Cronbach's alpha 58
13. Model fitting 59
14. Exposition 1: 59
15. Exposition 2 63
13. **Managerial Implications** **69**
14. **Future Research** **69**
15. **Conclusions** **69**
16. **Appendix** **70**
17. **References** **84**
18. **Summary** **92**

Introduction

This thesis analyzes deepfake technology from a more general perspective to a marketing point of view. Gives a possible prediction for its future and ends with quantitative research on the role of disclosure timing of such a technology could have in changing consumers' attitude towards advertisements, and the possible moderating effect of the terminology used.
The first part of the thesis begins with a brief deepfake history followed by a more technical analysis of the different types of it. Opportunities, risks, and possible solutions are carefully analyzed and discussed from both the general and the marketing perspective. The second part of the thesis is focused on developing the research questions, hypothesis, research methodology, and analysis. The research has been conducted with an ad-hoc built survey using Qualtrics and by exposing respondents to different scenarios from the combinations of the independent variable and moderator dimensions. The analysis of the gathered data has been done using RStudio and the full script can be found in the appendix. The data confirms part of the hypothesis. Disclosure timing does affect consumers' ad attitude but the terminology used doesn't seem to be responsible for any significant change.

History of Deepfake

In 2017, on Reddit, an unidentified user under the name of "deepfakes" uploaded some pornographic videos claiming they belonged to famous actresses such as Gal Gadot, Scarlett Johansson, Taylor Swift, Maisie Williams. The videos got labeled as fake but they were using a surprisingly accurate technology of face swapping based on deep learning. Reddit soon took out the videos, banned the user, and deleted r/deepfakes subreddit (Cole 2018) but the technology was so surprising that it quickly gained popularity. It started appearing in other subreddit, forums and got quickly noticed by news media like Vice that spotted the subreddit first (Cole 2017). Although the platform's attempts to ban the popular technique known as "deepfaking" and the follow-up action by other multimedia platforms like *Pornhub, Twitter, Gfycat,* and *Discord,* deepfakes kept spreading across the web like an unstoppable virus (Gardiner 2020).
The person who developed the deepfake was a software engineer and, alongside the deepfakes themselves, he released the development kit to allow any user to create their own manipulated videos efficiently. The technology used open-source tools available from big tech companies like Nvidia and Google that made everything accessible, free of charge, and available at any time with a little effort and knowledge needed. What that meant was that creating a deepfake became suddenly available to anyone (Gardiner 2020).
From that moment, a door was open that could never be closed.
Since then, the word deepfake has been used to describe "believable media generated by a deep neural network" (Mirsky and Lee 2021) and the world has seen an exponential growth of this type of content online.
The main difference between before and now is the fact that before these effects were only a prerogative of high skilled professionals who were working in the film industry or had at least some knowledge in visual effects. Now, instead, to obtain a believable result you only need a computer, an internet connection, and you are set to go.
The technology is evolving at a huge pace and it progressively requires less data and processing time to create one. In 2018 the app *FakeApp* required some consistent amount of input data to create one. In 2019, similar apps were already offering the same service but were less data demanding and more accessible. *Zao* for example, the Chinese app, only requires users to take a series of selfies with specific facial expressions and places the whoever face into a famous movie scene for free and in less than a minute (Kietzmann et al. 2020).
In January 2020, a new application called *Reface* by Neocortext inc. allowed users to do the same as these two previously mentioned applications, but only with one selfie and in roughly 30 seconds. [1]

The results are incredibly accurate. Even though it is still possible to spot some artifacts, with the aid of proper adjustment it's easy to trick the public into believing the content is real. An impressive result, yet shocking for the time, has been achieved in the BuzzFeed video published in 2018 where Jordan Peele made Obama say that former president (in charge at the time of the video) Donald Trump was "a total and complete dipshit".[2] The video aimed to warn about the risks of using such technology to mislead the public and create fake news. Of course, Obama never said those words, but the result looks incredibly real (Mack 2018).

Figure 1 | https://www.youtube.com/watch?v=cQ54GDm1eL0

From that video, the threat started to be clear to the big audience too. It was incredibly easy to create a video that could trick the mass into believing something that never happened.
The threat is so real, DARPA (Defense Advanced Research Projects Agency) [3] started two new projects devoted to Deepfake detection: Media Forensics (MediFor) in 2019 and Semantic Forensics (SemaFor) in 2020 (Sayler and Harris 2019).
Fortunately, even though there is still a big percentage of non-consensual deepfake porn around, the trend seemed to change and most of the deepfakes online today are user-generated and just for entertainment purposes or memes. (Johannes Tammekänd, John Thomas, and Kristjan Peterson 2020).

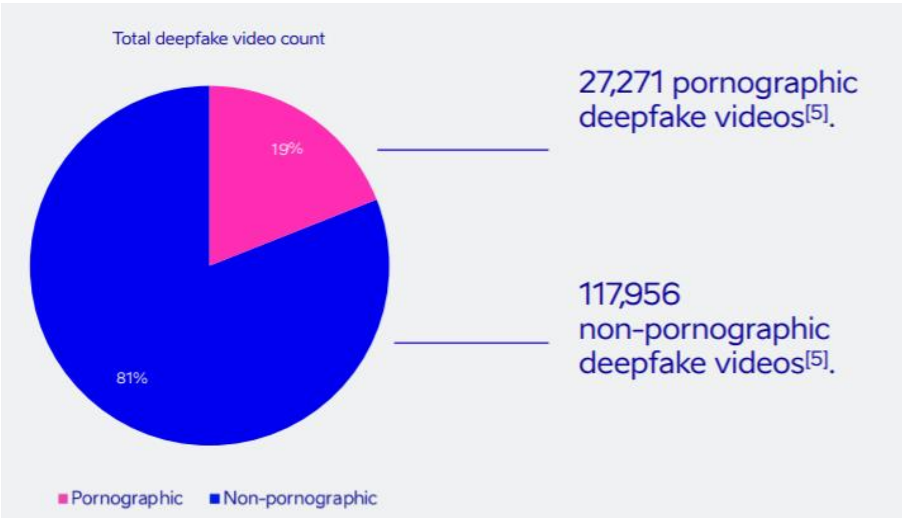


Figure 2 |https://thesentinel.ai/media/Deepfakes%202020:%20The%20Tipping%20Point,%20Sentinel.pdf

Some famous examples are the obsession with putting Nicolas Cage's face into every movie (Haysom 2018).

Figure 3 | https://geektech.me/wp-content/uploads/2021/03/a663e2c59f8665125a900f4f5f856ab7.jpg

The deepfake of Mark Zuckerberg “Imagine this...” [4] where he publicly announces how much power he has by having access to everybody’s data.

Figure 4 | https://www.instagram.com/p/ByaVigGFP2U/

And the latest, incredibly real @deeptomcruise's TikTok account by Chris Ume[5].

Figure 5 | https://cdn.mos.cms.futurecdn.net/vCiYBhEfxECz9ZEWX6YuQc-970-80.jpg.webp

All these can fall under the set of "funny videos and memes" but user-generated non-consensual porn and memes are not the only content the world has used deepfakes for.
Indian politician Manoj Tiwari, Bharatiya Janata Party (BJP) leader, in 2020 collaborated with a communication agency that used a deepfake to make the politician speak in 3 different languages, Hindi, English, and Haryanvi (an Indian dialect spoken by most of his electors).

Figure 6 | https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/

The Vice article (Christopher 2020) managed to obtain an interview with some BJP exponents. In the interview, they explain their goal was to reach as many people as possible in India.
Neelkant Bakshi, co-in-charge of social media and IT for BJP Delhi, told VICE[6] "Deepfake technology has helped us scale campaign efforts like never before" by "convincingly approach the target audience even if the candidate didn't speak the language of the voter" (Christopher 2020). This was the first time a deepfake has been used for an official campaign. The last time (at least recorded in this thesis) is the deepfake created by conservative Korean politician Yoon Suk-yeol's team where the politician answers questions to their potential electors (Moon Jae-in 2022).

Figure 7 | Screenshot of Yoon Suk-yeol deepfake

Deepfakes do not only consist of video or image manipulation but also audio. Similar to video deepfakes, the same concept can be applied to clone the voice of a target subject. A Text-To-Speech system can be created to generate speech audio starting from an input of just 5 seconds. It is also able to learn the tone of the subject accurately (Jia et al. 2018). The input file can then be used to train another Neural Network to puppet a target video and create a deepfake that doesn't need to use any impressionists to reproduce the voice of the target subject. It's all AI-generated and it needs very little skills and human interaction to be created (Thies et al. 2020).
Companies like Resemble.ai[7] or Descript[8] are already using these technologies to help creators augment their creative process by worrying less about minor mistakes. Any error can be easily fixed using text-to-speech synthesis and voice cloning allowing creators to focus more on the content itself.

The technology and applications have increased substantially since their exposition to the public in 2017. Data from thesentinel.ai report show that from 2019 deepfakes have been having a year-over-year growth of ~6820x reaching more than 100M+ of total deepfakes online in 2020 (Johannes Tammekänd, John Thomas, and Kristjan Peterson 2020).

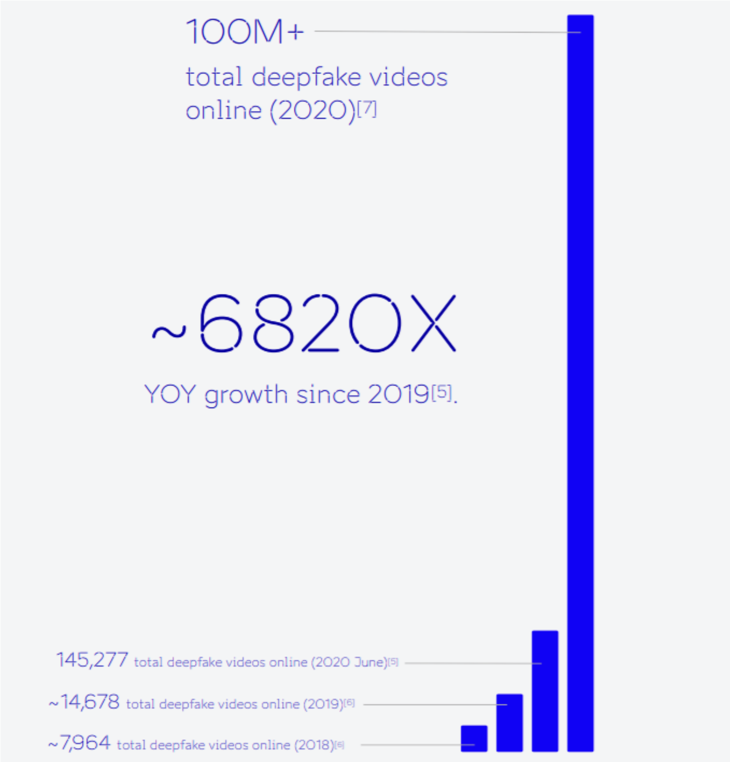


Figure 8 | https://thesentinel.ai/media/Deepfakes%202020:%20The%20Tipping%20Point,%20Sentinel.pdf

The number doesn't seem to stop its huge growth, especially after big tech companies like Snapchat, Apple, and ByteDance have integrated the function into their platforms (Dave Gershgorn 2020).
Deepfakes have entered the mainstream world (Heaven 2020). New companies are entering the market offering AI solutions and many like the already mentioned Descript, Resemble, Reface, etc. are exploiting deepfakes for creating new business opportunities. As of today, deepfakes still have some flaws and a trained eye can still spot some artifacts but the technology is in a constant high paced evolution and it's justified to think that in a couple of years it would be as nearly as impossible to tell, with the naked eye, the difference between real and deepfake content.
In the next paragraphs, we will explore deepfakes' threats, solutions, opportunities, and finally, the implications and applications for marketing.

Different types of deepfakes

The majority of the time deepfakes are thought of as the technique to superimpose the face of a source subject to a target subject, but it's not just that. Let's see together what are the different possibilities of this technique.

Facial Re-enactment

Face2Face

Facial re-enactment is the alteration of a target subject's facial expression in a video based on a source subject input. Face2face's approach was the first one to only require just monocular RGB input in real-time. The algorithm employs a dense photometric consistency measure to track the facial expressions of both the source and target videos. Reenactment is therefore accomplished through the rapid and effective transfer of deformation between source and target. The target sequence is used to obtain the mouth interior that best fits the re-targeted expression and warp it to produce an exact fit. Finally, it re-renders the synthesized target face on top of the associated video stream, blending it in with the real-world illumination. With this technique, it is possible to control, in real-time, the facial expression of the target subject that will follow what the RGB source will do (Thies et al. 2019). The figure below shows an example of this technique.

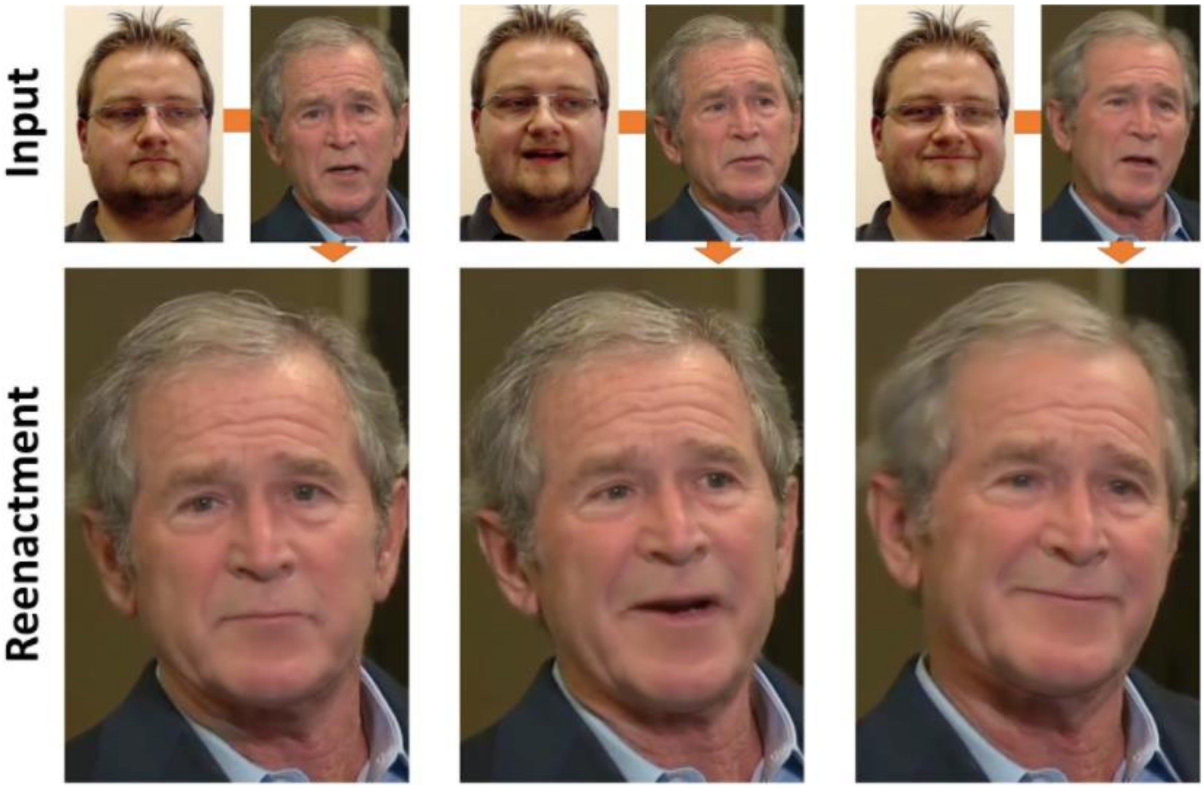


Figure 9 | Facial Re-enactment. Gorge W. Bush facial expression is controlled by another video source

Neural Voice Puppetry

Neural Voice Puppetry works following the same concept of face re-enactment, but instead of using video as a source, it uses an audio track. Starting from an audio sequence from a source human or digital assistant, the algorithm creates a photo-realistic output video of a target person that is in sync with the audio from the source input.
A deep neural network using a latent 3D face model space drives this audio-driven facial reconstruction. The model learns temporal stability inherently through 3D representation and uses neural rendering to generate photo-realistic output frames. This method is capable of using both voices of any unknown source or synthetic voices with text-to-speech algorithms (Thies et al. 2020).

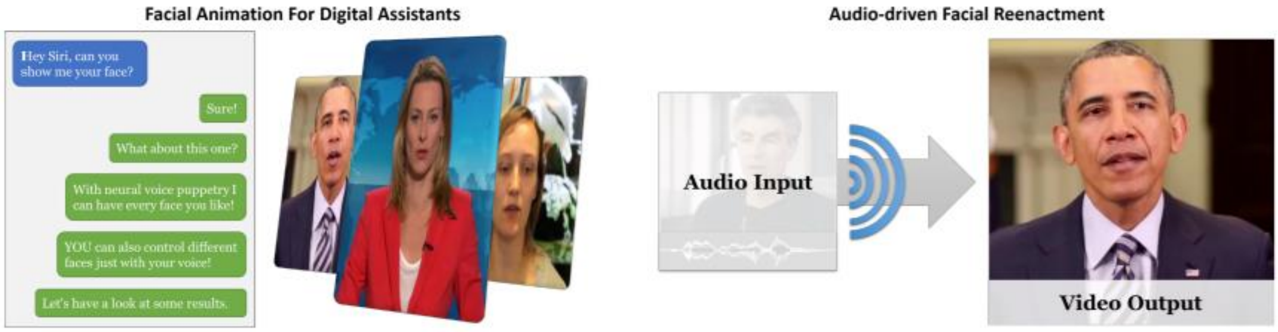


Figure 10 | Neural Voice Puppetry scheme

Face Swapping Deepfakes

This type of deepfake superimposes the face of a source subject to the one of a target subject. The algorithm analyzes both the target and the source face creating a latent space of similar features that are then used to perfectly swap the face of the source to the target in the most credible way. It's one of the most common uses of deepfakes (like the one used in the subreddit r/deepfakes) and it's the type of deepfake that it's mostly known to the public (Korshunova et al. 2017).

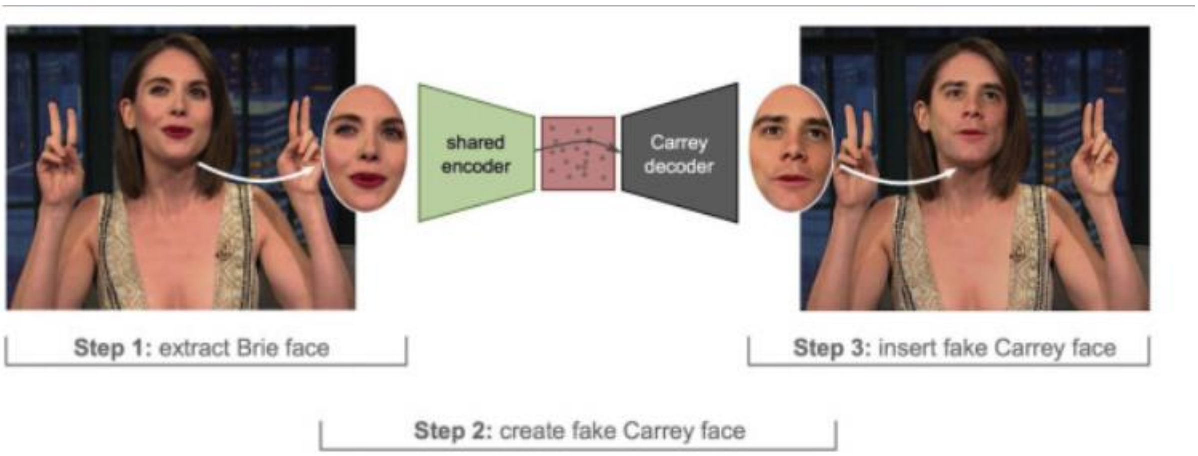


Figure 11 |Allison Brie original video (left), Jim Carrey deepfaked into Allison Brie (right)

Full Body Deepfakes

This type of deepfake expands the concept of facial re-enactment (Thies et al. 2019) to a full-body level by transferring the body movements of a source subject to the one of a target subject. Starting from a video of a person dancing, the algorithm can transfer the performance to a new target just after a couple of minutes of movements. The method employs video-to-video translation, with pose serving as an intermediary representation. It extracts poses from the source subject and applies the learned pose-to-appearance mapping to generate the target subject to transfer the motion (Chan et al. 2018).

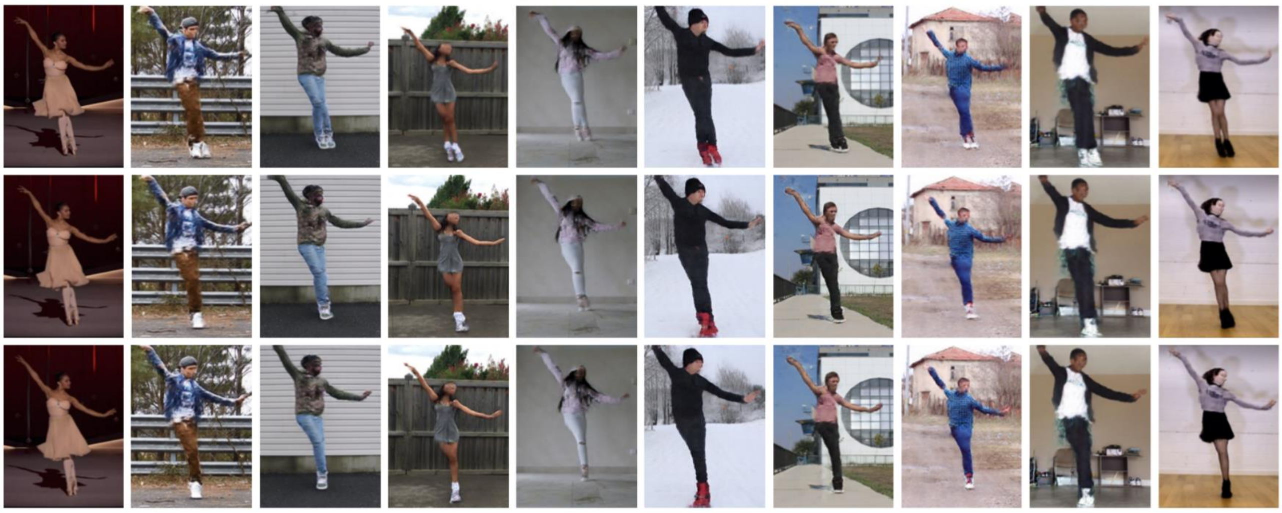


Figure 12 | Source video of a professional dancer (first column), superimposed motions (all the other columns)

Audio deepfakes

Deepfakes are not only image-based, they can also refer to generated audio that mimics a target. The AI can analyze a source subject from samples and it is then able to recreate the synthetized voice which is then possible to use in a text-to-speech system (Jia et al. 2018).

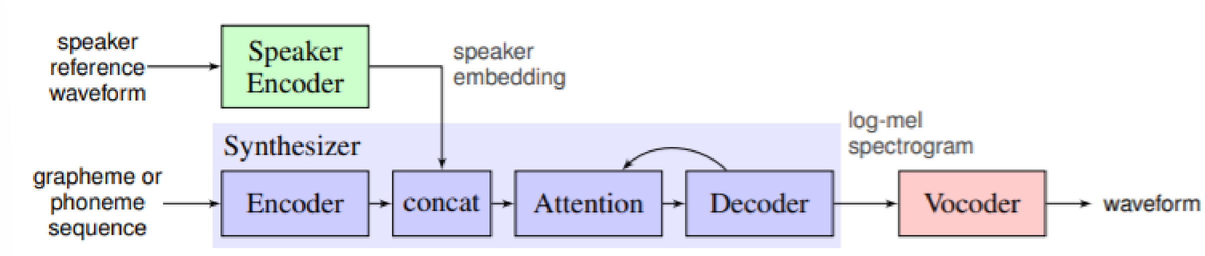


Figure 13 | Model overview of an audio deepfake

Concerns, threats, risks:

A lot of warnings have been raised by both the community and experts about the dangers of this technology. Whether it's philosophical, social, political, neuroscientific, psychological, every field seems to realize how impactful such a technology could be for the future society. A study by J. Hancock and J Bailenson (Hancock and Bailenson 2021) points out how deepfakes could impact the future and change our memories. It highlights how humans tend to give more importance to visual stimuli compared to other kinds and how we retain more visual information rather than verbal or written information.

Fake News

The first threat is the spreading of misinformation and fake news.
Fake news has been a major issue in recent years. Americans are more worried about the spread of misinformation than terrorism and violent crimes. They believe it will have a direct effect on the confidence in government, in each other's, and to the ability of politicians to get work done (Mitchell et al. 2019).

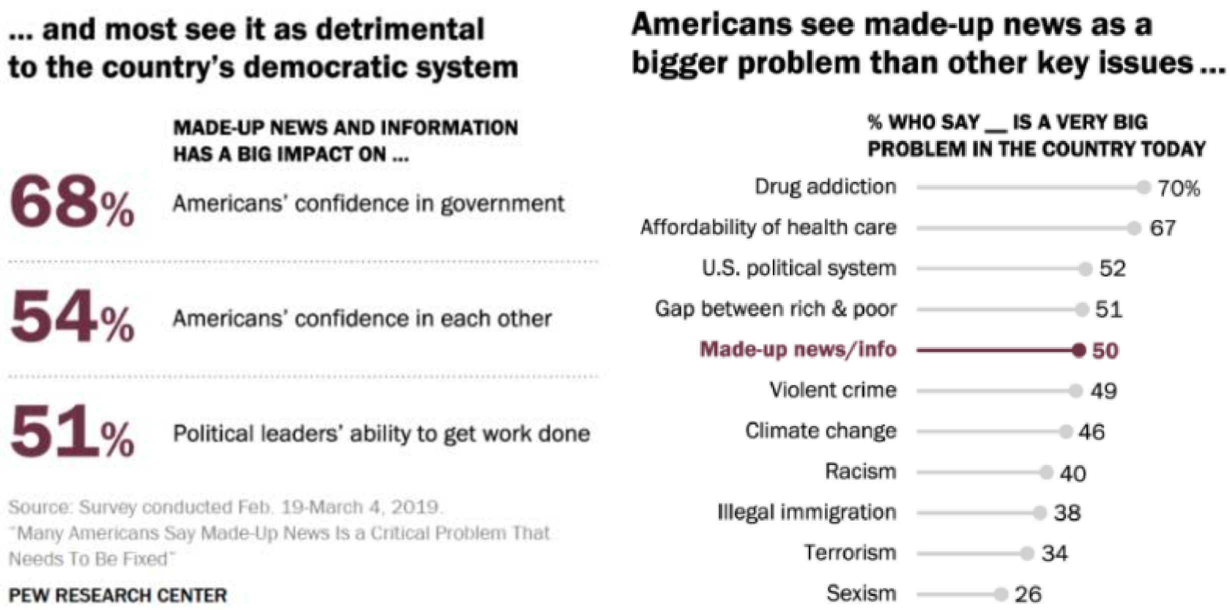


Figure 14 | https://www.statista.com/topics/3251/fake-news/#dossierKeyfigures

Simultaneously, the trust of news media is decreasing drastically. The last 5 years showed a reduction in trust in both national and local news has decreased from 76% to 58% and from 82% to 75% respectively.

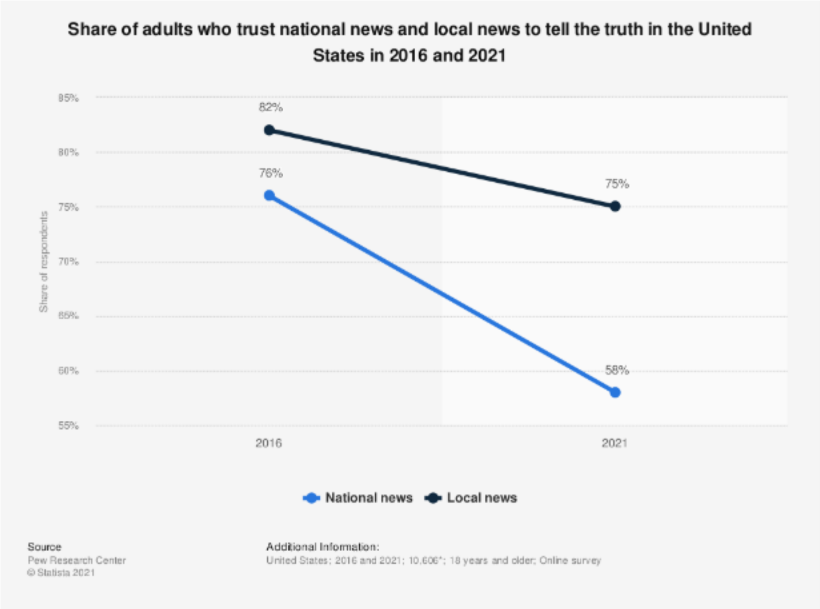


Figure 15 | https://www.statista.com/topics/3251/fake-news/#dossierKeyfigures

Adults in the US also showed a big switch in their level of confidence in recognizing fake news. (Amy Watson 2021)

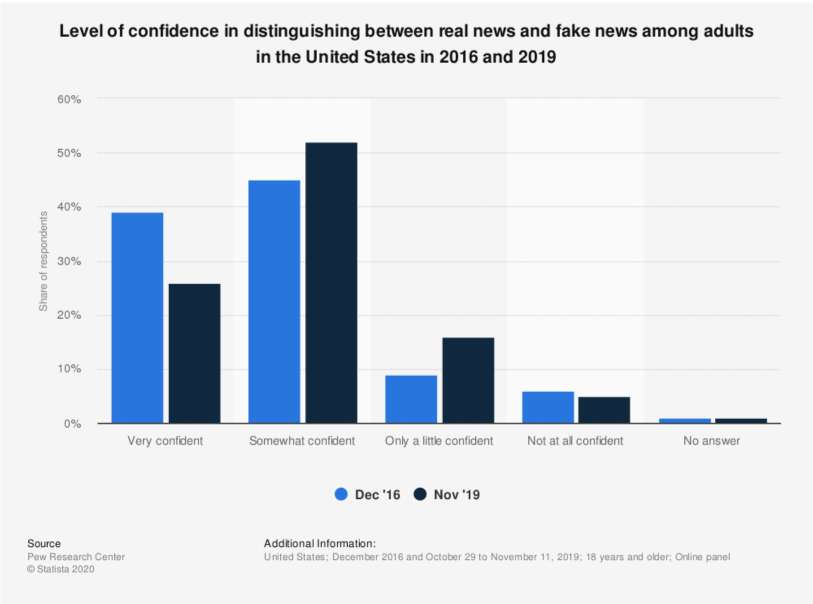


Figure 16 | https://www.statista.com/topics/3251/fake-news/#dossierKeyfigures

In this sense, deepfakes could be a big player in drastically further dropping the trust of the public towards news organizations.
In 2019 WhoIsHostingThis.com surveyed 980 Americans. It emerged "82.1% of people think there will be more distrust in information in the next 10 years" and when exposed to the deepfake technology 88.8% of the respondents said, "deepfakes would cause more harm than good" (Johannes Tammekänd, John Thomas, and Kristjan Peterson 2020).

Democracy

If used improperly, deepfakes could lead to mass manipulation, election interfering, political uncertainty, and they could potentially undermine democracy. Eventually, they could also turn into a problem of national security (Westerlund 2019).
“Putting words in someone's mouth on a video that goes viral is a powerful weapon in today’s disinformation wars, as such altered videos can easily skew voter opinion” (Westerlund 2019).
In 2018 Vooruit [9], a Belgian Socialist party, used a deepfake video of Donald Trump asking the Belgians to follow America’s lead and exit the Paris climate agreement [10].
Even though the video presented a lot of artifacts and the fakeness is pretty clear, some people still believed the video was real (Von Der Burchard 2018).

Figure 17 | https://www.facebook.com/Vlaamse.socialisten/videos/10155618434657151/

A study published in late 2020 proved how microtargeted deepfakes could negatively affect the attitudes of citizens towards a politician and his party (Dobber et al. 2021).
The application of this into, for example, a Facebook targeted ad, could lead to serious consequences.
The before mentioned study points out how some groups of people might be more sensitive to this type of miscommunication. Christians could (as the study suggests), for example, amplify the effect of the misinformation if compared to the same type of misinformation but not targeted to a specific group (Dobber et al. 2021).

Some tailor-made deepfakes could be the spark of some attacks or other serious consequences for the population safety if used in a specific delicate time, targeting the right group of extremists,
In summer 2020 in Cameroon, one year before the upcoming elections, a video of Christophe Guilhou, ambassador of France, appeared online. The video was a deepfake of him saying "The French Republic, is the supervisory power that colonized Cameroon". It quickly spread but France24[11] debunked it as quickly as it spread (Holubowicz 2020). This was a perfect example of a microtargeted deepfake.

Figure 18 | https://journalism.design/le-deepfake-de-christophe-guilhou-embarrasse-au-cameroun/

In Gabon in 2018. After Ali Bongo Odimba, president of Gabon, disappeared from the public eye for some months, the news he might be severely ill started to rise as well as some tension between the different political parties. To try and calm the public opinion, the president suddenly appeared on Gabon 24 official tv[12], giving the traditional speech of the end of the year. The video created a lot of debate on whether it was real or fake. The president looks unnatural, he rarely blinks his eyes, and the look on his face looks fake. Adjembe Etogho, member and spokesperson for the Gabonese Council of Resistance, said some members of the opposition tried to show and prove the video was a deepfake. The information the video could be fake or staged started spreading quickly. In January 2019, Gabon's elite Republican Guard seized the national radio station. Lt. Kelly Ondo Obiang called the Gabonese to stand up and "restore democracy". The latter has been found later to be real by running it through two different deepfake checking algorithms. The first one was run by Steve Grobma, CTO at McAfee which scored 92% real. The second test was run by Siwei Lyu digital-media forensics expert which scored 0.99 (0 = fake, 1 = real). The video and the misinformation behind it sparked an attempted coup. (Cahlan 2020)

Figure 19 | https://www.facebook.com/watch/?ref=external&v=324528215059254

Pornfakes and cyberbullism

Not only politics and democracy can be harmed by deepfakes, but also individuals.
We already mentioned the high incidence of *pornfakes*, used to substitute the victim's face into a porn actor/actress's body to create nonconsensual adult content. Usually, as thesentinel.ai reports, the majority of the target are celebrities or people of influence (Johannes Tammekänd, John Thomas, and Kristjan Peterson 2020). These videos mostly end up in either pornographic websites or underground communities, for example, on Telegram[13]. Some channels offer a customed pornographic deepfake service to users who are willing to pay (Johannes Tammekänd, John Thomas, and Kristjan Peterson 2020).
Cyberbullying, Revenge Porn, Blackmailing, and *Ransomfake Attacks* using pornfakes as a weapon are already a thing. The perpetrators ask for cryptocurrencies payments threatening the victim to release the pornographic deepfake to the public.
An incredible story is the one that happened to 18-years-old Noelle Martin. Everything starts when she randomly decided to do a reverse google image search of herself. The result was hundreds of explicit images of her face photoshopped into the bodies of porn actresses. When she dared to speak out, other perpetrators started bullying her using images of her public appearances to create more pornographic deepfake videos that got uploaded into various porn websites (Melville 2019).

Scams and impersonifications

The next big tool for scams will be voice cloning. We already mentioned some companies that help creators edit their video and audio files using video deepfakes and audio deepfakes like Resemble.ai[14] or Descript[15]. The technique is very useful if used properly but at the same time, voice cloning could be used by scammers for impersonating another person. This already happened twice. The first time to a CEO of a UK-based company in 2019 where the scammers managed to impersonate his boss, the CTO of a parent company. The scammers ordered him to transfer €220.000 to the bank account of a Hungarian supplier using a cloned voice (Jesse Damiani 2019). The second time the same happened, to another person from a different company, with the same technique, but for a much larger amount of money: $35 Million (Brewster 2021).

The Liar's Dividend

Besides the creation of fake realities, another type of distortion that could arise from the advancement in technology of deepfakes is what Chesney and Citron refer to as the *Liar's dividend* (Chesney and Citron 2018).
The Liar's Dividend is a term used by Chesney and Citron to describe the emerging and growing tendency of denying reality by claiming a certain video or audio was altered using deepfake technology (or more generally manipulated) or using the same in favor by creating new manipulated content that will be used as evidence against the claims.
Their theory is aligned with the trends we've seen before. The more the public learns about the existence of these types of manipulations; the more they lose trust in what they see; the more is going to be easier for liars to escape from their responsibilities.
"Put simply: a skeptical public will be primed to doubt the authenticity of real audio and video evidence. This skepticism can be invoked just as well against authentic as against adulterated content." (Chesney and Citron 2018).

A crystal clear example of what was described happened in June 2020. Winnie Heartstrong, a Republican who was running for congress started sharing conspiracy theories claiming George Floyd died in 2016, and the videos of his death on May 25, 2020, were all created using deepfake technology.

Figure 20 | https://www.dailydot.com/debug/republican-candidate-george-floyd-conspiracy/?utm_medium=email&_hsmi=91059372&_hsenc=p2ANqtz--osmn58tvLTZD2M1jH4xfl6125WtgAjALoHdY2wSTXSGvDjWinhskL5wSPkbRnKW0qUxwR6TCGPqpPfeZnmTmS5gIX-qGRp3a2oONOTO5aiHDysuo&utm_content=9

Heartstrong also created a website called investigatedeepfakefloyd.com where she further claims the deepfake was created by using images of Floyd and former NBA star Stephen Jackson. She also claimed how former Police Officer Derek Chauvin was Cash Cab host and comedian Ben Bailey.[16] (Thalen 2020)

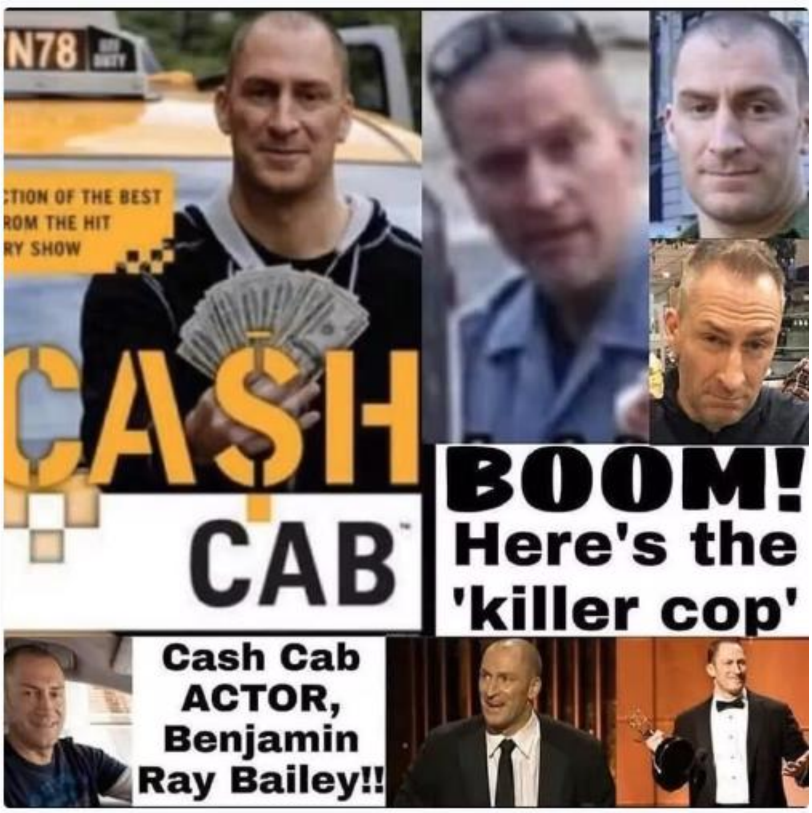


Figure 21 | https://investigatedeepfakefloyd.com/

As Dailydot.com reports, Henry Ajder, Head of Threat Intelligence at deepfake-monitoring service Deeptrace labs, said how this story is "one of the most divisive attempts to weaponize the idea of deepfakes" he has seen. "It reinforces the concerns we've repeatedly raised that once awareness of deepfakes has spread sufficiently, the idea will be seized on by fringe groups and malicious actors not just to dismiss fake things as real, but also to dismiss real things as fake," Ajder told the Daily Dot. "These cases illustrate how for a worrying number of people, deepfakes provide the missing thread of explainability that suddenly make conspiracy theories appear much more justifiable." (Thalen 2020).

Visual Dominance

The rise of deepfakes is wiping away what we used to take as granted: "Seeing is believing".
The visual dominance and importance in respect to other signals is well known since 1976 (Posner, Nissen, and Klein 1976) and a study by Koppen and Spence in 2007 shows how visual information is the one humans rely most on (Koppen and Spence 2007).
Even if the results did not show any gigantic difference in creating false memories compared to text input, they did prove that deepfakes can be responsible for creating them (Murphy and Flynn 2021).
The technology still isn't perfect, and some flaws and artifacts are common but to cite a catchphrase by Károly Zsolnai-Fehér's famous youtube channel *Two Minutes Papers* [17] [18]: "Huge improvements are just two papers down the line".

Don Fallis analyses the *epistemic threat* of deepfakes (Fallis 2021). In his study, the philosopher explains how videos in later years have become the "gold standard" of truth. The ability to easily manipulate videos could change this standard and the distrust born of deepfakes would prevent people to believe what they saw.

Solutions and countermeasures to deepfakes:

A technical overview and discussion.

Detection algorithms

The counterattack against deepfake is not an easy one and people are not sure whether it should be tech companies' responsibilities, the government's, news organizations, or other institutions.
The best solution would be a combination of all these. Institutions and news organizations need to inform and help the general public to become more aware and literate about these new technologies. At the same time, the governments need to update their legislation on what are the boundaries of fair use of deepfakes and include the new generation of cybercrimes.
Lastly, it's essential to make more research and studies on technologies that could precisely detect synthetic media. In this sense, DARPA (Defense Advanced Research Projects Agency) [19] started two new projects devoted to Automated Multimodal Media Manipulation detection: Media Forensics (MediFor) in 2019 and Semantic Forensics (SemaFor) in 2020 (Sayler and Harris 2019) gathering a top team of researchers led by Accenture Federal Services (AFS), Google/Carahsoft, New York University (NYU), NVIDIA, and Systems & Technology Research. The aim is to create systems that could improve deepfake detection (DARPA 2021).
Alongside these projects, universities, and researchers from all over the world are trying to develop the best detection algorithm. The problem though is to end up in a mouse vs cat game. Every time an algorithm becomes better at detecting a deepfake, the deepfake algorithm becomes better at creating them (Kietzmann et al. 2020) (Lomtadze 2019).
Table 1 shows what are the different methods for detecting deepfakes using algorithms (Yu et al. 2021).

Table 1 | https://www.koreascience.or.kr/article/JAKO202125761199587.pdf

Methods	Description
General network based methods	CNNs accomplish the task of detection, which is a frame-level classification task.
Temporal consistency based methods	Due to flaws in the forgery algorithm, anomalies between neighboring frames are discovered in deepfake videos. As a result, RNN is used to detect such inconsistencies.
Visual artefacts based methods	Underlying image disparities in the blending boundaries would result from the blending operation in the generation process. These artefacts are identified using CNN-based approaches.
Camera fingerprints based methods	Devices leave varied footprints in collected photos due to different generation processes. Faces and background images are simultaneously detected as coming from separate devices. As a result, these traces can be used to complete the detecting task.
Biological signals based methods	Concealed biological signals in faces are difficult to interpret using GAN, making it impossible to synthesis human looks with appropriate behavior. Biological signals are extracted based on this observation to detect deepfake films.

In the FaceForensic++ paper (Rossler et al. 2019) researchers created a new detection pipeline using the face tracking method by Thies et al. to extrapolate the region of the video where the face is first (Thies et al. 2019). This improved the performance of the detector. The extrapolated area then goes through different classification methods. Generic manipulation detection (Bayar and Stamm 2016) (Cozzolino, Poggi, and Verdoliva 2017), computer-generated vs natural image detection

(Rahmouni et al. 2017), face tempering detection (Afchar et al. 2019), and XceptionNet are used for detection. The latter outperforms all the other methods (Chollet 2017).

Figure 22 | https://arxiv.org/pdf/1901.08971.pdf

To train the detectors, the researchers created a large-scale dataset of manipulated media based on both classical methods like *Face2Face* (Thies et al. 2019) and *FaceSwap* (Kowalski 2018) and learning-based approaches like *DeepFakes* ("Deepfakes/Faceswap: Deepfakes Software For All" 2020) and *NeuralTextures* (Thies, Zollhöfer, and Nießner 2019).
They aim to create a detector that could be applied in a real-world scenario, so considering different codecs and video quality.
The study gives a forgery detection reference point for humans by surveying 204 participants (mostly computer science students, so in a way, literate about the technology) where they had, in a time limit between 2, 4, or 6 seconds, decide whether the image was real or fake. The ratio was 50:50 between real and fake images. The results show the performance of the different technologies in tricking the respondents and the results are also divided into the quality of the image (RAW, HQ, LQ).

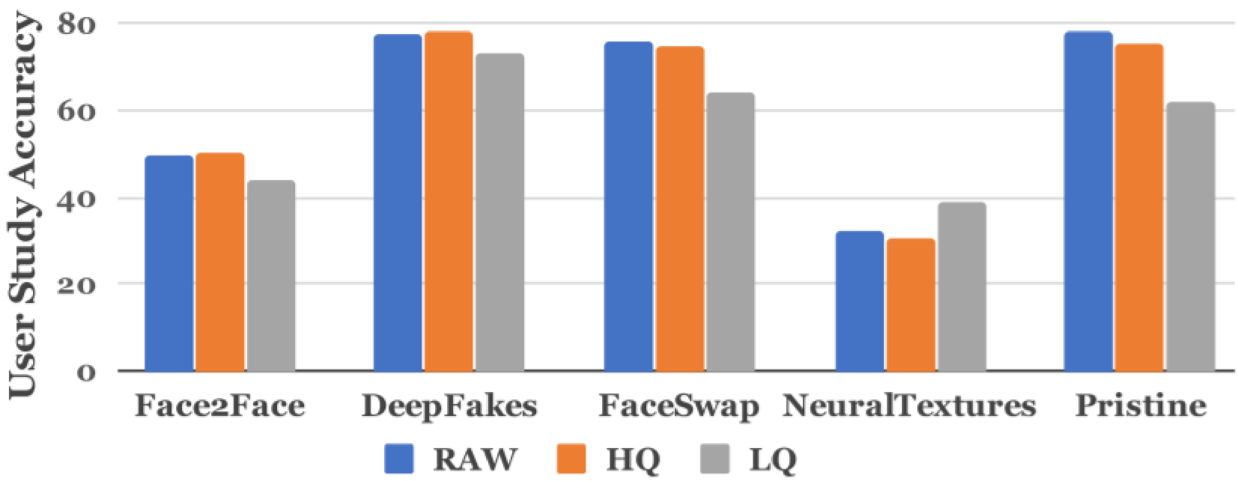


Figure 23 | https://arxiv.org/pdf/1901.08971.pdf

Results highlight the incredible superiority of the algorithm compared to humans.
Figure 24 shows the results for the different algorithms used and the different qualities and methods of forgery (Rossler et al. 2019):

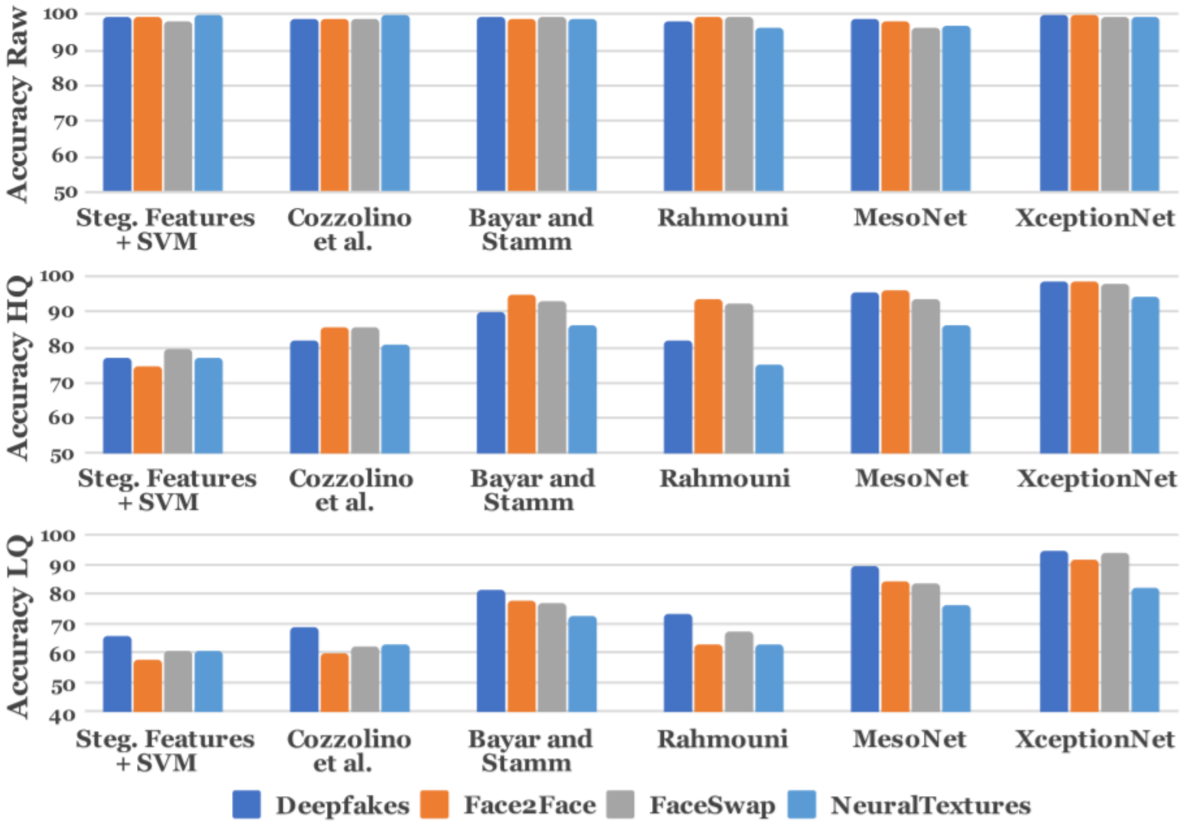


Figure 24 | https://arxiv.org/pdf/1901.08971.pdf

At the link http://kaldir.vc.in.tum.de/faceforensics_benchmark/, it is possible to see a benchmark for different custom algorithms. The benchmark auto-updates to make it easier for researchers to compare results.
As mentioned, though, as the technology advances; the detection algorithms become less precise and vice-versa resulting in a never-ending chase.
Some other studies suggest instead using humans as a *collective intelligence* by aggregating their results. The study shows that when done, humans are as accurate as an AI (the one used in their experiment at least). Another solution suggested by the same paper was a combination of AI and humans as an augmentation for human deepfake detection. Humans can better contextualize other variables that go beyond the mere video analysis and are better at detecting contextual flaws while AI can focus on detailed image analysis. (Groh et al. 2022).
Considering the weaknesses algorithms could have in detecting deepfakes, some suggest a different point of view for solving the problem: the *blockchain*.

Blockchain architecture

”Blockchain is a peer-to-peer immutable network that consists of an ordered sequence of linked and duplicated data blocks.” (Lee 2021) where its integrity is maintained through public-key encryption for each network interaction and update (Drescher 2017).
The blockchain is based on a distributed system and a decentralized immutable consensus process that is safeguarded by encryption protocols that control each block in the chain.
The degree of centralization is the primary differentiator between a blockchain and a centralized database. While all database entries are centralized, each blockchain participant has a safe copy of all records and alterations, allowing any user to see the data's origin. Figure 25 shows the architectural differences between the two systems.

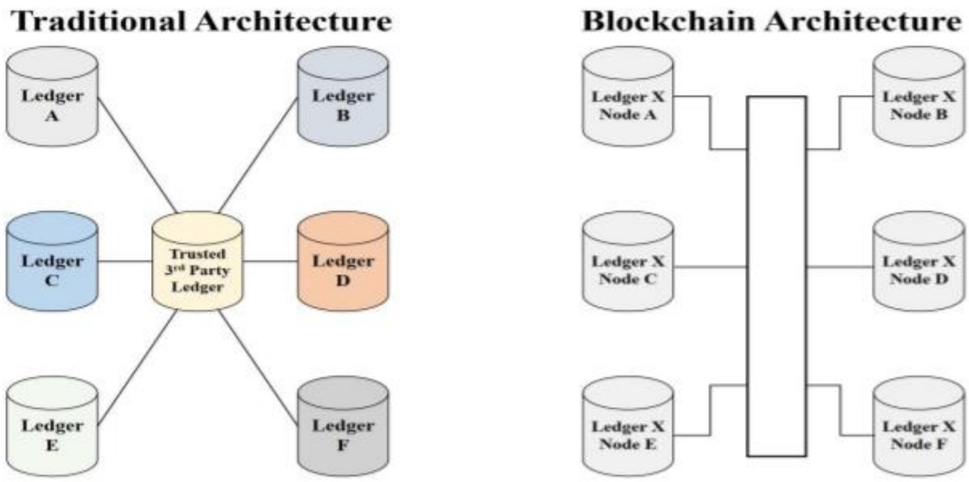


Figure 25 | https://www.koreascience.or.kr/article/JAKO202125761199587.pdf

“The previous block’s hash is combined with the current block’s hashed content and a timestamp to create the new block.” (Lee 2021). In this way, every block will always be dependent on the prior block and the 2 will be closely linked. By replacing a block in the blockchain, the chain is torn apart and a new blockchain is created. Every block that is not interchanged will be no longer valid.
The only way to modify the blockchain is for the majority of the system's nodes to accept the new blockchain. (Hemlin Billström and Huss 2017)

Smart Contracts

Smart contracts are self-executing contracts in which the conditions of a multi-party agreement are written directly in code. A blockchain network connects the code and the agreements contained inside it.
Smart contracts eliminate the need for a central authority, legal system, or external enforcement mechanism to carry out reliable transactions and engagements between remote, anonymous participants. As a result, transactions are traceable, transparent, and irreversible. Due to the inherent challenges of smart contract generation, developers have invented domain-specific languages such as Solidity[20] to make the process easier (Mohanta, Panda, and Jena 2018).

Hyperledger Fabric

The Hyperledger Fabric framework is a scalable architecture-based framework for distributed ledger systems that offer high degrees of confidentiality, robustness, flexibility, and scalability. Like other blockchain technologies, it consists of a ledger, smart contracts, and some method for members to keep track of their transactions. Members enroll using a trusted Membership Service Provider (MSP) rather than an open permissionless model that allows anyone to join a Hyperledger Fabric network (requiring transaction validation and network security mechanisms such as "proof of work"). It also enables users to build pathways, which enable a group of businesses to form their transaction ledger. (Hyperledger 2020).

Why the blockchain?

When a user with a digital identity wants to do anything, they may be requested to prove their identity before being granted access to something (such as finances or movies). Blockchain, with its tamperproof records, logs, and transactions, may be used to establish the validity and originality of digital media in a decentralized, trusted, and secure manner, making it the greatest alternative for countering deepfakes. Many academics have already suggested *Smart Contracts* and *Hyperledger Fabric*-based systems to be used to regulate and record the history of digital content transfers. This could assist in content authentication and preventing deepfakes spread.
IBM built a blockchain based on Hyperledger Fabric and offers solutions to countering deepfakes and fake news (Chaban 2020).
Safe.press[21], uses IBM solutions to check sources and track authenticity. A green *safe.press stamp* is placed on each page where a member publishes a press release or article. The stamp acts as a digital seal of approval that is tied to a blockchain key.

Figure 26 | https://safe.press/

That key is instantly registered on a blockchain ledger constructed by Block.Expert using IBM's open-source Hyperledger Fabric. When one of these registered news sources is attached to future stories or references, the key associated with it is tracked (Chaban 2020).

Blockchain authenticates content

By 2023, up to 30% of world news and video content will be authenticated as real by blockchain, countering deep fake technology.

Although fake news has existed for centuries, social media bots have rapidly accelerated the rate at which this deliberate disinformation can be spread. In addition to traditional news stories, technology is being used to create convincing fake audio and video. However, organizations and governments are now turning to technology to help counter fake news, for example, by using **blockchain technology** to authenticate news photographs and video, as the technology creates an immutable and shared record of content that ideally is viewable to consumers.

Figure 27 | https://www.gartner.com/smarterwithgartner/gartner-top-strategic-predictions-for-2020-and-beyond

Content authentication will be soon fundamental for maintaining a high level of trust from the public and avoiding consequences driven by the new technologies. Gartner's 2020 predictions highlight that by 2023 30% of the world news will be authenticated by the blockchain to counter deepfakes (Panetta 2019).

The future of deepfakes seems to be in danger, but while in some areas deepfakes are dangerous, for some others they are a godsend.

Opportunities: when deepfakes are a godsend

This paragraph will be the UNO reverse card[22] for deepfakes. In this part of the analysis, we will explore potential benefits and opportunities for this technology. We will go through already existing implementations and possible ones for the future.
Knives are dangerous, they could cause harm to people and lead to deaths and hospitalizations. If used properly though, they could help you cut some nice vegetables for making a delicious dish. Sometimes, it just depends on the use people do of some tools. The same applies to deepfakes. If used properly, deepfakes could enhance and augment the experience of people in many ways, from art to marketing, therapy, and many more industries.

Edutainment

The Dalì Museum in St. Petersburg, Florida[23], features a life-size talking deepfake avatar of Salvador Dalì created using historical archival footage. The museum allows its customers to interact with the painter. It’s possible to ask him questions, and he would directly explain his pieces of art. At the end of the journey, funnily and engagingly, he asks people to take a selfie using a camera placed on top of the little totem that will be later sent to the customer (Mihailova 2021).

Figure 28 | Dalì deepfake's kiosk

This application falls into what it’s more broadly called *edutainment*. Deepfakes can enrich art and education by providing new audio and visual content. Deepfakes could assist educators and professors in delivering more engaging courses than traditional visual and media.
AI-generated synthetic media can bring historical personalities to life in the classroom, making it more engaging and interactive.
Using Microsoft Hololens[24] or other AR/VR tech, it could be possible to create a mixed or virtual reality where students could approach history classes in a completely new and engaging way while, for example, interacting with historical figures of that time (Jaiman 2020)
JFK's resolution to stop the cold war speech was recreated cloning JFK’s real voice and speaking style. The speech was recreated by CereProc by using a Deep Neural Network and Unit selection voices over eight weeks where the CereProc team analyzed recordings of 831 speeches and constructed the voice by dividing them into 116,777 tiny phonetic components. Chris Pidcock, co-founder of CereProc explained how “the unit selection voices reproduce the character of the speaker and the DNN voices help guide the prosody modeling which is the intonation” (CereProc 2018). The same can be applied to recreating short movies or documentaries starring the interested character.

Accessibility

An interesting and life-changing application of deepfakes is the creation of audio-text-to-speech technologies that could help people who lost their voice to finally have it back. It would be possible to completely reconstruct someone’s voice by using archived content, videos, voice messages. A startup called Project Revoice[25] tries to achieve this. The project helps people who lost their voice due to medical conditions such as motor neuron disease to get it back. A voice-cloning technique tries to recreate the voice of the patient with the same accent and intonation. VOCALiD[26] uses voicebanks and proprietary voice blending technology to generate distinct vocal personalities for any device that converts text to speech for those who have speech and hearing problems.

Therapy and mental health

A combination of deepfake technology and immersive XR technologies could boost the perceived level of realism and serve patients who are trying to recover from a certain disease or trying to do the rehabilitation. A study after another has found that when a person learns a skill in virtual reality (VR), the expertise frequently transfers to the physical world. The question of whether this effect could be boosted by more personalization arises. If we could simply take a patient not just to a beach, but to their beach, complete with unique features and landmarks and creating an incredibly accurate copy of themselves. How much more effective could this type of experience become as technology develops? For treatment reasons, VR allows a doctor to customize the therapy session, allowing them to give stimuli tailored to each patient in a safe and private environment. AR and VR also promote immersion and a sense of presence (the sensation of "being there" in a virtual world), giving the experience a sense of immediacy. Finally, VR shares the mechanism of embodied simulations with the human brain. To put it another way, both the brain and virtual reality rely on simulations of the physical body in the real world to represent and predict behaviors, concepts, and emotions. As a result, it is possible to create specific virtual environments that can imitate both the external and inside world/body (Wiederhold 2021).

The Dutch documentary *Deepfake Therapy* depicts how this technology can be deployed to assist those who are grieving the loss of a loved one. People can now have realistic video discussions with a deepfake of someone who has died while being overseen by a skilled therapist. Patients with post-traumatic stress disorder (PTSD) are frequently treated using exposure therapy, which entails encouraging them to confront and conquer their phobias by exposing themselves to the things or events they dread in a secure atmosphere. Deepfakes could improve exposure therapy by recreating synthetic content that depicts a specific person or event that the patient is afraid of (KRO-NCRV 2020).

Removing Language Barriers

Google has been working a lot on audio cloning and synthesis. They recently published a new paper on version 2 of Translatotron. “Translatotron 2 is a neural direct speech-to-speech translation model that can be trained end-to-end. Translatotron 2 consists of a speech encoder, a phoneme decoder, a mel-spectrogram synthesizer, and an attention module that connects all the previous three components.” (Jia et al. 2021). It uses technology from another Google Inc. project from a couple of years before called Tacotron 2, a neural network architecture for speech synthesis from text (Shen et al. 2018). We will soon be able to speak another language, in real-time, using our own voice.

Face re-enactment as well can come in handy when having to translate movies or, for example, during a foreign speech. The interpreter's voice can be blended into the target subject’s face that will mimic the interpreter’s lips (Vincent 2021) (Prajwal et al. 2020). Flawless.ai is a startup that does just this. They focus on lip-syncing visualization in different languages [27]. An example of this in the act is the social intervention campaign Malaria Must Die[28], in which former English player David Beckham seems to speak nine different languages in a vocal petition to stop malaria.

Freedom of speech

Going in an opposite direction to what was exposed in the concerns and risks paragraph, this technology could be used to preserve freedom and help human rights activists and journalists to stay anonymous in dictatorial and oppressive regimes. For citizen journalists and activists, using technology to report injustices on traditional or social media may be incredibly empowering. Deepfake can be used to mask voices and faces to safeguard their privacy.

Personal digital avatars provide autonomy and can assist individuals in expanding their purpose, thoughts, and beliefs, as well as enabling self-expression, which may be challenging for some (Chesney and Citron 2018). HBO[29]'s Welcome To Chechnya documentary applies this concept. Director David France goes to Chechnya to speak with a group of human rights activists confronting the anti-LGBTQ+ persecution in the Russian Republic of Chechnya. The documentary uses deepfake technology instead of blurring to anonymize the people that appear in the documentary without losing the facial expressions and emotions behind the stories (Rothkopf 2020).

Figure 29 | top left: doc's frame, top right: volunteer lending her face for the documentary, bottom left: capturing different angles of a face, bottom right: frame from the documentary

Film Industry

The film industry is certainly the one that could benefit the most from deepfakes. Youtube channel Shamook[30] showed in a video how Netflix[31] could have exploited *DeepFaceLab*[32] opensource free software to de-age their actors in *The Irishman* instead of spending money on expensive CGI effects that drove the budget of the movie as high as $US175 Million (Chilton 2020). The result has been described by some fans as “mind-blowingly better” than the original.

Figure 30 | Comparison by Shamook's Youtube Channel

The possibilities in this field are unlimited. It could be possible, for example, to resurrect dead characters. It’s theoretically possible to have Marilyn Monroe acting in a 2022 movie with a tenth of the budget and the time needed for recreating her using CGI, and probably with a better effect. This is the case of Spanish beer Cruzcampo[33] which managed to bring iconic Lola Flores[34] back to life with the use of deepfake for their *Con Mucho Acento* campaign (Vega 2021).

Needless to say the same can be just applied to actors who are still alive. A deepfake of a famous character will be way less expensive for production companies than flying him/her out, paying for hospitality, and having them act. Actors in the future might just have a dataset of their faces ready to send to companies who require the use of it (Kahn 2021).

The creator of many famous deepfakes like the already mentioned @deeptomcruise page on TikTok, Chris Ume, has started a new company called *Metaphysics.ai*[35] they aim to sell deepfake solutions to both the film and videogame industries.

When it comes to post-production and reshoots, the benefits of deepfake technology for the business can be massive. Filmmakers could replace lines without having to reshoot the whole scene. A new line could be just rewritten, and the scene will be automatically fixed. Instead of having a full set and crew to fix minor changes, a tiny team might do the same with a computer and in a fraction of time.

Gaming industry

Another industry that will surely benefit from such technology is that of video games. Applications could vary from deepfaking yourself into videogames to needing less time to create virtual characters. A combination of FaceSwapping techniques (Kowalski 2018) pose estimation (Xie et al. 2021) and full body deepfakes (Chan et al. 2018) could allow us to perfectly transfer our face and personal types of body movements into a video game character.
Chris Ume, from his Instagram's[36] page (@vfxchrisume), shows an example of how the famous football game FIFA[37] could look like if EA Sports[38] used deepfakes instead of CGI. The results are insane (D'Angelo 2021) (Ume 2021).
NetEase[39], a giant Chinese gaming company, uses AI to automatically create in-game characters from a single picture. The difference from other deepfakes is the combination with a 3D bone-driven model which predicts facial parameters with a big physical significance (Shi et al. 2019).

Figure 31 | Left CGI, Right Deepfake

A big deal for marketing

Marketing-wise deepfakes could be a huge game-changer. Besides the already mentioned Dalì museum in Florida, this technique could help marketers create deeper bonds and higher engagement with their customers if leveraging the right levers. Companies like Zalando[40] could make it possible for customers to deepfake themselves into their e-commerce platform to see how garments look on them (Whittaker et al. 2020).
Zalando already used deepfake technology in 2018 with the #whereveryouare campaign that managed to create bespoke messages to 12 countries with 290,000 local versions for small villages and towns in Europe. The campaign gained more than 180M impressions across social media and saw an increase of 54% in-shop sales and a 79Mln increased revenue (Infinitizer 2018).
The fashion industry could probably benefit more than the others from this technology. Companies like *Artificial Talent*[41] are offering GANs solutions to generate models for promotional purposes. The models can be customized following the needed physical and facial characteristics the company requires. The generated characters can then be used for their brand image or to create promotional campaigns without having to cast any human model or having to pay for a photoshoot (Whittaker et al. 2020).
Balenciaga's[42] SS22 fashion show titled *Clones* was a virtual runaway featuring cloned avatars that were CG-scanned onto deepfake models. The music of the digital event was an AI-generated version of *La Vie En Rose*. It was the first, all AI-generated fashion show and a marketing success (Ahmed 2021).

Figure 32 | Balenciaga's SS22 show

Virtual Influencers

Whittaker et al. suggest brands could generate synthetic influencers or endorsers to perfectly represent their brand and aesthetic. An existing famous example is Lil Miquela, built using CGI. With 3.1 Million Followers Lil Miquela has endorsed brands like Samsung, Prada, Calvin Klein, and Mini (Powers 2019).

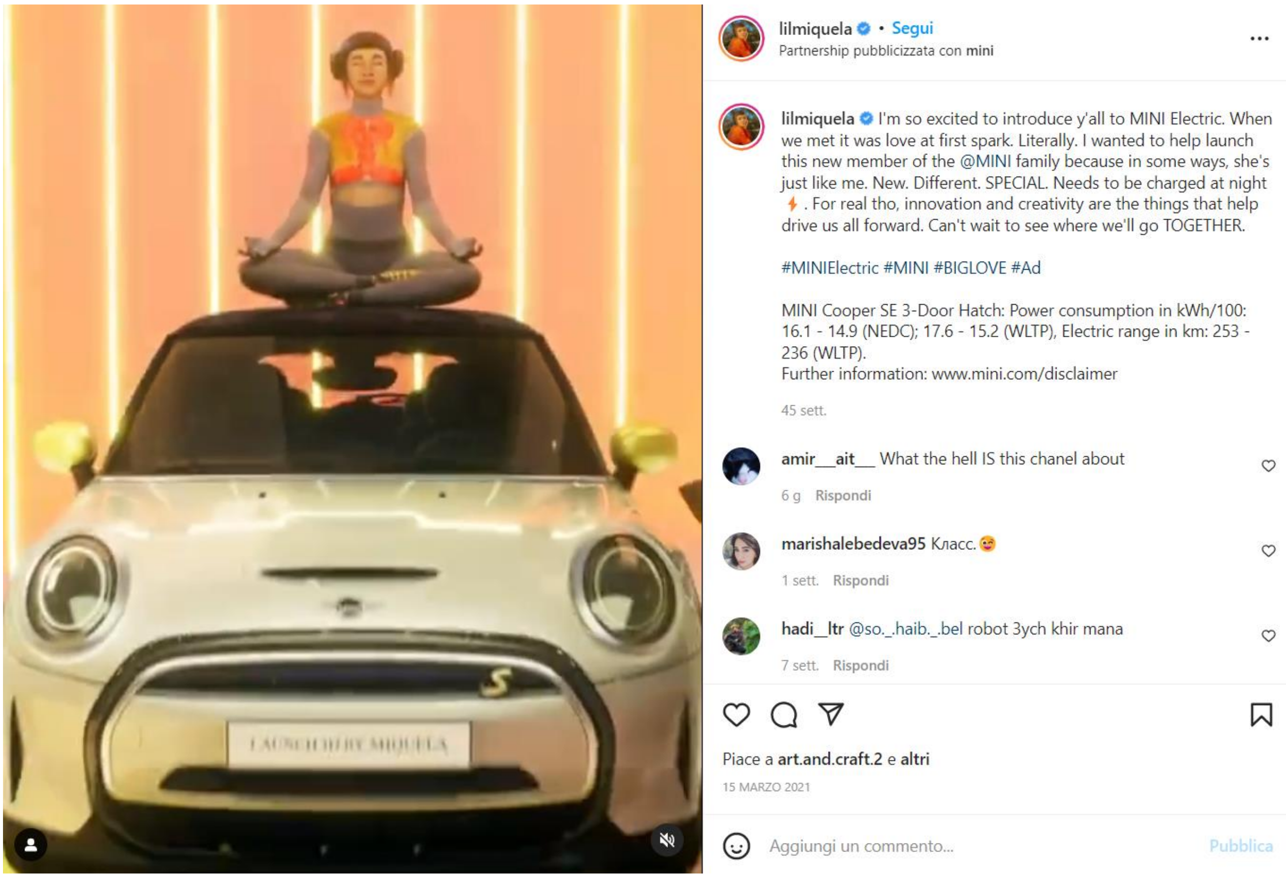


Figure 33 | Lil Miquela X Mini

Synthetic endorsers can have the brand create their narrative and life drama, all without the unforeseen dangers or concerns that a human endorser can introduce (Koh & Wells, 2018; Powers, 2019). Synthetic endorsers retain the controlled elements of traditional spokescharacters, but with higher potential to express human qualities similar to those of the target audience. This results in a greater sense of authenticity and relatability. Life story, beliefs, personality, and physical appeal, can all be programmed into a synthetic endorser to produce a perfect message source for the brand's audience. Synthetic endorsers can feel authentic because an individual's love or admiration of a fictional figure can be linked to increased perceived realness (Koh and Wells 2018).

A recent study interviewed 30 virtual influencers' (VI) followers to understand their perspectives and found some interesting. Followers are frequently content-driven, eager for fun, and suspicious of phony brand promotion on social media. Some are unconcerned about whether this is met by humans or virtual characters. Because of intriguing story-driven innovative content, VIs create natural interest and maintain fans. Furthermore, VIs that are very humanlike are more appealing to followers, demonstrating a low level of *Uncanny Valley* effect, signaling acceptance (Choudhry et al. 2022).

The Uncanny Valley Effect

The uncanny valley effect is an unsettling feeling of unfamiliarity individuals experience when viewing or interacting with robots that are nearly but not quite identical to humans. Chatbots that use synthesized voices or reproduce human faces but not perfectly have been shown to produce this effect that negatively impacts the experience of interaction (Ciechanowski et al. 2019).
Deepfakes could overcome this problem by creating a more authentic and realistic avatar, enabling them to have a stronger ability to imitate emotions, hence improving perceived emotional intelligence and reacting to human needs for empathy and understanding in real-time. Facial expressions play an important role in social interaction coordination, acting as a visual tool that helps the message recipient grasp the communicator's emotions, ideas, and intentions (Keltner and Haidt 1999). Positive emotions such as enthusiasm, humor, and attachment may help weak persuasive messages to be accepted more readily (Griskevicius, Shiota, and Neufeld 2010). Brands could deploy them to deliver promotional messages, help the customer during the shopping experience, or even become a proper virtual assistant like the one proposed by Samsung’s Neon[43] (Hitti 2020) (Whittaker, Letheren, and Mulcahy 2021).

Figure 34 | Samsung's NEON avatars

Bigger reach

Similar to the Zalando #whereveryouare campaign and the Malaria Must Die campaign featuring David Beckham, deepfakes can help reach a broader audience and have different versions of the same ads from only one take.
In 2020, Synthesia[44] partnered with ad agency Craftww[45] for JustEat’s[46] commercial featuring Snoop Dogg[47]. The American rapper performs a bespoke song about the company’s service.
The only problem was they could not use the same ad in Australia, as the service there is known as Menulog[48]. The solution was genius and saved JustEat a lot of money. Another version of the ad was created using deepfake technology that both changed the lyrics and lip movements in every single moment of the commercial without having to take extra shots (Whitaker 2021). A combination of some of the technology analyzed before could potentially create a worldwide ad from only one take. The impact this could have on time and cost of production is massive.

Hyper-personalized ads

If Zalando's example is what is closer to a real case of hyper-personalization. We could imagine beyond that and hypothesize that, in the future, it could be possible to create even more hyper-personalized ads. Imagine you open your smartphone; you open the YouTube app and choose a video. An ad starts, but instead of seeing a random actor playing, you see a person that reflects your characteristics, or, even more, yourself. This is already possible with current technologies and could be the future of what ads might be in the future. People could have their deepfake avatars and the biometrical information could be used by companies to create specifically hyper-tailored ads based on the data available. More easily and less scary, brands could create apps that use deepfake technologies allowing customers to deepfake themselves into an already made advertisement allowing them to share it to their social network (Whittaker et al. 2020). In this way, customers can become co-creators of value in the advertising, shifting from being just a passive audience. The user not only becomes part of the ad and feels more engaged, but could also actively share the hyper-personalized ad starring him/her/they to social media platforms, helping the company to reach more potential quality customers (Kietzmann, Mills, and Plangger 2021). A study from 2017 suggests how there's a positive correlation between increased personalization and ad's persuasiveness and perceived benefit (Ham 2017) which "could increase the desired (re)action of the consumer" (Kietzmann, Mills, and Plangger 2021). Hyper-personalization and co-creation have been already tested out by a couple of brand pioneers like Frito-Lay's[49] campaign starring footballer Lionel Messi. An app called *Messi Messages*[50] allows customers to customize a personal message in 10 different languages and choose from thousands of names available. The tool provides a template message that can be edited with something similar to a mix between a chatbot and a drop-down menu to change the recipient's name, the sender's name, and the style of the message which is an invitation to watch a match or play an online game (Kulp 2021).

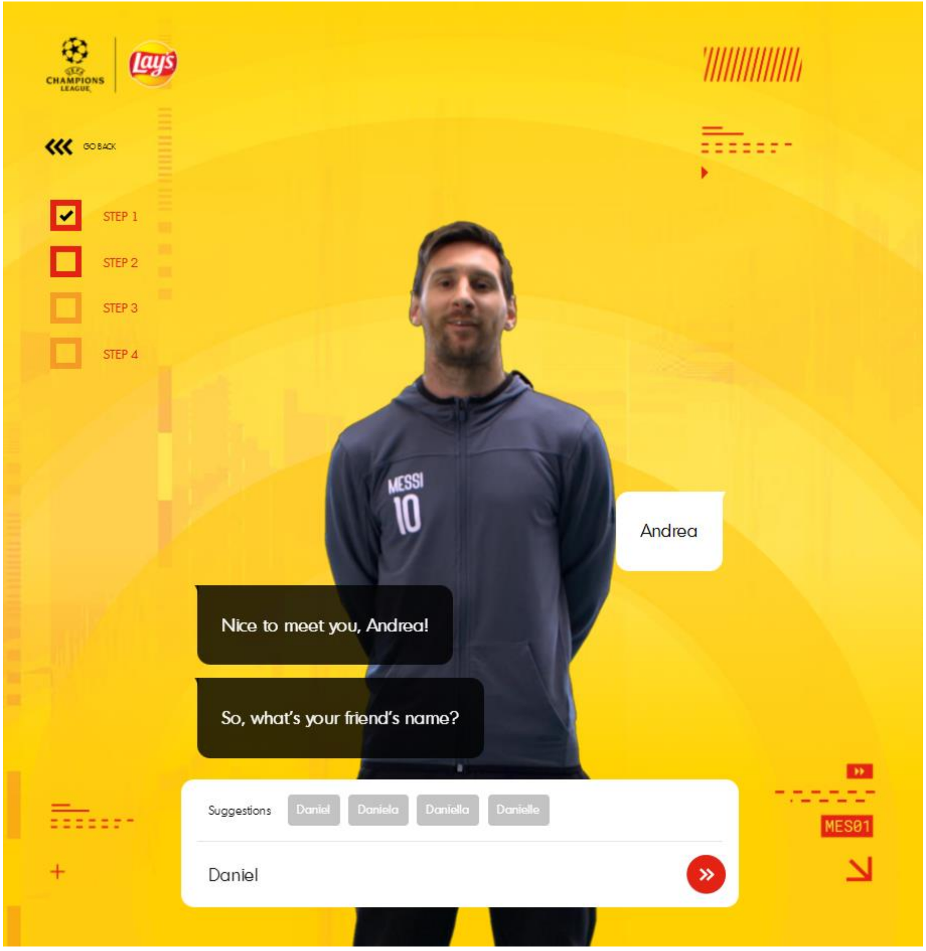


Figure 35 | Frito Lay's Messi Message

Spotify, in 2020, considering the huge success of Spotify Wrapped (Cury 2021), and singer The Weeknd, decided to push the limits of Wrapped's concept by creating a one-to-one experience with the Canadian singer. The experience is called *Alone With Me* and uses user's data and geolocation to understand what time of the day it is and, similarly to wrapped, historical data on past listens to create a super tailored message. The experience ends with a glitched video and the reproduction of some songs from After Hours' album (Leung 2020).

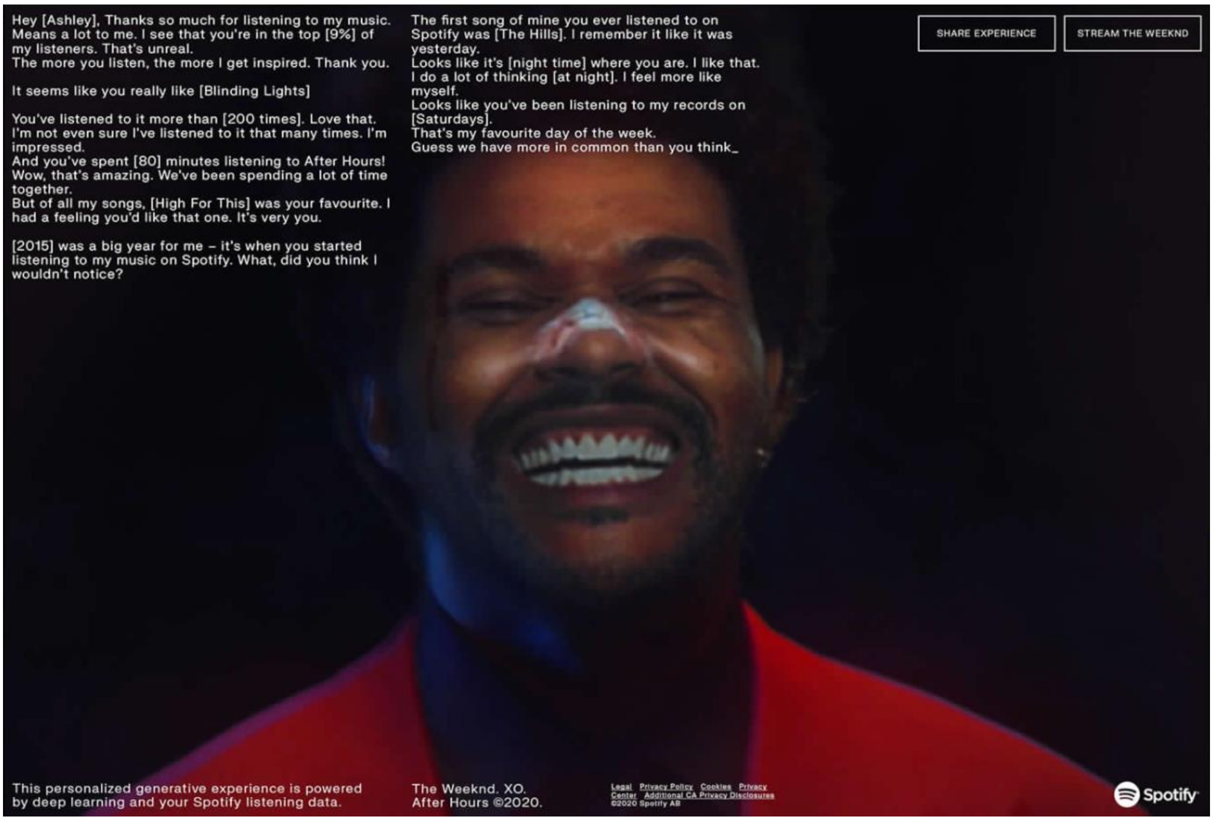


Figure 36 | Spotify's Alone With Me

Doritos's[51] #CoolRanchDance campaign for 2020's Super Bowl is a collaboration with the deepfake app called *Sway*[52]*.* The *Magic Dance* function creates full-body deepfakes based on their commercial showing a dance battle between Lil Nas X and Sam Elliot over Cool Ranch tortilla chips. Users must hold their iPhones stable for 30 seconds while filming their movements. The app then transposes the video over the body of a professional dancer. The result is a video that seems like the user is dancing perfectly to the choreography (Williams 2020). Co-creation has shown an increase in customer loyalty and overall value (Cossío-Silva et al. 2016). It would be interesting for future researchers to try and address this question specifically applied to hyper-personalization and deepfakes.

Tourism

Tourism could benefit from this technology too. Travel agencies could create super-resolution images of the destination and offer customers to be deepfaked in the environment. VR headsets could be implemented to give a little taste of what they would look like in the destination before going. The digital experience can be shared with your friends, partner, or travel mates. This would enhance the experience and trigger the activation of the customer (Kwok and Koh 2021). Deepfakes enable customers to create a realistic visualization of their consumption experience and improve the ability to imagine how they would use the product (Mulcahy et al. 2019) resulting in a higher purchase intention.

The Metaverse

The metaverse is probably going to be one of the big next things after Facebook announced its launch at the annual facebook AR/VR conference in late 2021 (Paul 2021). The metaverse is a parallel universe, a physical replica of the web, accessible via a virtual reality headset, and with its independent economy enabled by blockchain.

The metaverse concept is founded on three basic components.

1, Virtual Reality: This alternate universe will be accessible via augmented reality tools like virtual reality glasses or VR headsets.
2, Blockchain: the metaverse's economy is based on decentralized data transfer technology, which allows the production of virtual money and the trading of NFTs.
3, Web 3.0: the metaverse is seen as the next generation of the web 2.0 that we know today. The concept of Web 3.0 is centered on the concept of owning "pieces" of the Internet.
The metaverse will also enable the creation of one of the most popular digital items at the moment: NFTs. Non-Fungible Tokens are cryptographic identifying keys that enable the possession of any digital asset, such as an image, video, or audio file.

In the metaverse, the marketing opportunities are infinite. Brands could have virtual synthetic generated influencers to represent the brand with which their customers could interact and engage more directly and interactively (Csiszat, Szilagyi, and Milievic Renata 2022). It is possible, that one day we could have a philosophical conversation with Apple or Coca-Cola's generated avatars on the metaverse.

Risks for Marketing

Before we learned a knife could cause both harm and good. We learned it's better to use our knife to cut us some fresh food rather than harming others, but we must be careful to hold it from the right side. Marketers need to hold the knife from the handle and not from the blade.
This paragraph analyzes the possible risks of deepfakes implementations in marketing.

Data use and privacy

Besides the risk of fraud or fake news about the company that might circulate over the internet and that could result in a direct monetary loss or image damage (Helms and Mutter 2020), the first concern companies and marketing agencies should be aware of is the importance of transparency about the use of personal data; whether it's being stored, for how long, in which servers, what type of data, and what's the purpose of using and storing them.
Consumers feel vulnerable when they realize their personal information has been collected without specific consent for creating personalized ads; resulting in drastic drops in metrics like click-through-rate (CTR). Conversely, the same CTR increases when data are instead collected overtly (Aguirre et al. 2015). Companies should put more effort into building trust with their customers by clearly disclosing everything about the collection and use of their information. This specific phenomenon has been seen to be even more pronounced for smaller brands (Mukherjee, Smith, and Turri 2018).
Usually, customers are faced with what's referred to as *adversing calculus*. Adversing Calculus is the weight scale customers use between benefits (e.g., hyper-personalization) and costs (e.g., privacy). It's unclear how the scale works and what is the curve that defines when the maximum point of personalization accepted is reached before the perceived benefit for the customer stops being worth the cost in privacy. It's not easy for companies to find the perfect balance between marketing intel, building and maintaining trust, customer surveillance, and improving customer satisfaction.

In addition to transparency, though, brands could show the value gained by the customer if they share certain data, inducing the customer to actively make the final choice (Plangger and Watson 2015).
In a denouncing article from 2019, The Washington Post[53] shared concerns about the use the famous FaceApp does of its users' data and uploaded pictures. "Looking under the hood of FaceApp with the tools from my iPhone test, I found it sharing information about my phone with Facebook and Google AdMob, which probably help it place ads and check the performance of its ads. The most unsettling part was how much data FaceApp was sending to its servers after which … who knows what happens. It's not just your face that FaceApp might gobble up — if you age friends or family members, their face gets uploaded, too." The journalist then continues the article by reaching out to the CEO of the company Yaroslav Goncharov who explains some parts of their terms of use of personal data, trying to reassure their users on the topic (Fowler 2019). Similarly, the same happened in a Forbes[54] article, followed by another one just the day after (and another one by CNN), exposing the viral Chinese app ZAO by raising "concerns about privacy and safety issues" about the app request of too many rights over user's pictures and personal data than what is necessary to run the service (Doffman 2019a) (Doffman 2019b) (He, Guy, and Wang 2019). This should teach companies how it's very easy for them to risk their reputation and consequently their revenues if not enough transparent.

Perceived reality

The International Journal of Advertising, in a study by Kietzmann et al. makes a difference between truthfulness and truthiness. "Truthiness is a consumer's attribution of validity based on how something seems or feels to them, independent from and regardless of its truthfulness" (Berthon and Pitt 2018). The paper continues by analyzing if truthiness in this post-fact, full of deepfake world could eventually erase some of the concepts we've learned about like the *confirmation bias*, *selective exposure, desirability bias, repetition effect and priming*.
The previously mentioned Liar's Dividend (Chesney and Citron 2018), for example, could impact the perception of reality, confusing people that might not be able to recognize what's real and what is fake anymore. The *Sleeper Effect* (Hovland and Weiss 1951), states that even if consumers are aware that they have been exposed to fake content, the persuasive effects of that content have a lasting influence on later perceptions, even if the credibility of the content is shown to be questionable or even misleading. Even though the study is from 1951, we could probably apply the same funding to AI-generated synthetic ads. No matter if consumers are skeptical, the messages in those advertisements could be fundamentally more convincing in the long term. Deepfakes have the specific characteristics of being misleading and hyper-real. This could cause a shift in what we've been knowing so far in marketing. These are just hypotheses and we cannot have any proper answer until the use of deepfakes becomes more prominent (Kietzmann, Mills, and Plangger 2021).

Ad falsity perception and verisimilitude

According to the framework proposed by Campbel et al., *Ad Falsity* can be divided into 3 subcategories. False product-related claims, product-unrelated claims, and presented reality.
False product-related claims are the ones that falsely state something about the characteristics of the product or its effect. The misleading information can be either presented through text (e.g., "this product will make you lose 20kg in 10 days") or visually (e.g., by showing a before/after picture that is unlikely to be true). These types of false ads have been a concern for policy makers that made illegal misleading ads in Europe (EUR-Lex 2006) or, with a mandatory disclosure (in the US) (Hastak and Mazis 2011) (Hoy and Stankey 1993).
The second one, product unrelated claims, are when an ad's statements about a product or service are technically correct, but it may contain other inaccurate information (Petty and Craig Andrews 2008). False information is often excluded from the concept of deceptive advertising since it is assumed it may be easily discovered by consumers or have little impact on market behavior. An ad showing President Lincoln endorsing a smartphone, for example, will be obviously false.

Lastly, the presented reality of an advertisement can be false if it shows a version of the world that is different from what actually is, or if the world is generated artificially. This type of falsity can happen in multiple ways, especially in advertisements. For example, the techniques used to make food appear fresher, tasty, and perfect in supermarkets or restaurants' ads are well known (Little 2014). Green screens, stuntmen, and especially photoshop and CGI can be used to augment reality with animation or by eliminating imperfections/editing reality. These methods have always been used but lately, a new gray area is emerging, and some countries already got it covered. The beauty industry and the glorification of hyper-beauty and perfection have led to many negative consequences to adolescent girls' self-esteem already, and manipulated images contribute even more to the idealization of the perfect body/beauty, but with the little cherry on top which is they are fake (Kleemans et al. 2018). For this reason, many countries, starting from Israel, began to make mandatory a disclaimer saying the image was manipulated (Krawitz 2014) (Borau and Nepomuceno 2019).

Figure 37 | Before Photoshop (sx), After Photoshop (dx)

The main problem with ad falsity is the fact consumers are well known to negatively react to false information in advertisements. This comes from a tendency of self-protection or coping with persuasion attempts (Friestad and Wright 1994). The cited study from Campbell et al. suggests and links other studies on fakery in general that could be easily applied to how consumers would perceive deepfakes in ads (C. Campbell et al. 2021). It's possible, though, that the association is not as direct; the topic should be studied more in-deep with specific regards to deepfakes and differentiating in their different use too. The technology is so different from what we have seen so far it needs an ad-hoc study to properly understand the different perceptions of consumers.
Gaining a deeper understanding of how awareness of ad falsehood influences customer behavior has become critical as Campbell suggests. Research made so far on persuasion knowledge reveals that if consumers are aware that an advertisement has been modified, they may react differently to it (Friestad and Wright 1994). What could also impact the effect on the consumer can be the type of product, In the clothing industry, consumers might find the deepfake technology to have more benefits than costs.

Government policies

As mentioned in the previous paragraph, a grey area of ad manipulation is emerging, and like the photoshop example, the same could happen to deepfakes. China in 2019 made deepfakes without a disclosure a criminal offense just a month after California did a similar thing banning them during the election season (Statt 2019)(AB-730 2019). Facebook and Tiktok already banned deepfakes from their platforms (Shead 2020) (Statt 2020) and are investing in deepfake detections. At the same time, Google and Microsoft are known to be working towards deepfake detection systems (Kelion 2020)(Dufour Nick and Gully Andrew 2019). The mentioned DARPA programs are working as well and funding the research towards developing better detection systems (DARPA 2021). The world seems to prepare itself for the inevitable coming of deepfakes to the mainstream world and it's plausible more and more countries and governments will join the fight against them. Companies need to be aware of that before jumping too soon into investing an inconsiderate amount of money into this new thing which could be, eventually, completely banned from the mainstream. What could be more probable is that it will be mandatory to disclaim its use. It's of vital importance to understand what the best way is to disclaim the technology, how to do it, and when. The research proposed later in this thesis addresses this specific question.

Consumers' reaction to low budget ads

A study by Ducoffe tells us consumers make assumptions about how much a company has invested in advertising (Ducoffe 1995) and many others say consumers usually respond better to an advertisement that appears to be more expensive both in terms of creativity and placement (Modig, Dahlén, and Colliander 2014)(Lange, Rosengren, and Blom 2016)(Dahlén, Granlund, and Grenros 2009). Consumers weigh the perceived cost of an advertisement against the cost imposed on them in terms of time or effort. They react badly when they consider an ad to impose a comparatively high cost on them relative to the advertiser's costs (e.g., spam.)(M. C. Campbell 1995). Consumers could react negatively if they realize the cost of producing a deepfake ad is too little. At the same time though, we've seen big brands like Balenciaga using it recently (Ahmed 2021) and having the opposite effect compared to what has been hypothesized by the mentioned studies. Further research is of extreme importance to understand if deepfakes could be the exemption about perceived budget importance or if, being new, it's just a temporary effect due to the innovation of it.

General perception

There is one last reason why deepfakes might not be a successful move for companies, and it's pretty simple: Maybe people just don't like deepfakes and that's it. Maybe the general fear, concerns, and stories revolving around deepfakes (e.g., porn, or fake news) just make deepfakes not a "good thing" in the eyes of the consumer. This is something marketers should understand and be aware of. However, the few studies that have been conducted seemed to show a relatively positive attitude towards the technology by the users (Poon 2021). To add a more interesting fact to this, a study by Wu, Ma, and Zhang, shows us how deepfakes can sometimes have an expected *Self-enhancement effect* on people who are exposed to deepfakes of themselves into celebrities bodies. The studied subjects felt more confident and beautiful after being exposed to deepfakes, highlighting some new possibilities and frontiers in treating body images disturbances (Wu, Ma, and Zhang 2021).
One thing that will be addressed later in this research is if the word "Deepfake" itself has an intrinsic bad connotation and perception in the mainstream. By changing the terminology, for example from "deepfake" to "synthetic ad", marketers could modulate the perception of consumers about the technology.

Research Question

Considering the discussed framework and the position deepfakes have in public opinion where they are seen as dangerous and a tool for creating more harm than good. It is reasonable to assume the grey area deepfakes are “living in” will soon become a more defined and colorful one. We’ve seen how some countries are already taking care of the issue and started creating laws that ban or regulate more in deep the use of it and the rest of the world is believed to be following them. We covered the important role deepfakes could have for marketing and industries like cinema or videogames. The technology could have a big impact in terms of cost, time, and creativity. Brands, managers, and the literature must study this phenomenon more in detail and all its shapes. Most of the papers found during the literature research were screaming in help for academics to cover some aspects of the topic. This research tries to help cover what is probably going to be the soonest problem that will need an answer. Considering how countries have been moving against deepfakes, it is understandably right to think it will be soon mandatory to disclose the use of this technology to the public. This research wants to help understand how disclosure could impact consumers’ perception towards the advertising that uses deepfakes and tries to study what is the best way to do it. One study in 2014 revealed how disclosure timing impacts the persuasion of an ad. The study analyzes sponsorship disclosure and its effects. It shows that when the disclosure is placed before, or during the promotion, it negatively impacts the persuasion effect because the consumer processes the information differently becoming more critical and creating a stronger resistance to the information conveyed. (Boerman, van Reijmersdal, and Neijens 2014). This is in line with other studies found that similarly stated that when a consumer knows that they are being manipulated, or somebody is trying to manipulate them they respond negatively and they are more aware of the advertising calculus (M. C. Campbell and Kirmani 2000) (M. C. Campbell 1995) (Darke and Ritchie 2007). In the deepfake world though, we still don’t know if this could be easily applied. It could be possible that stating the use of deepfakes beforehand could lead to more trust or just being prepared and not feeling tricked by the ad or the brand. It’s for this reason that the first two research questions are:

Q1: *Does deepfake disclosure affect consumers’ ad perception?*
Q2: *How does deepfake disclosure timing affect consumers’ ad perception?*

The study also wants to analyze if the term used for defining the technology could moderate the effect of disclosure timing. Deepfake is a term that has been widely used in the colloquial world and is usually used in the mainstream when talking about fake news, pornfakes, ransomfake, manipulated elections, threats for democracy, misinformation, and disinformation in general. It is right to assume that the word deepfake triggers (intentionally, and unintentionally) a sequence of links to bad emotions. Conversely, terms like synthetic media, AI-generated media, generative media, are more technical but still do describe the same set of techniques and, the study theorizes, to be less linked to a disruptive dystopic future. For practical use, “synthetic media” has been chosen for the research. The third research question then is:

Q3: *Does using “synthetic media” instead of “deepfake” when disclosing moderate the effect of it?*

Arguably, some people might not know what deepfake or synthetic media mean. For this reason, after exposing respondents to the first stimuli, they are being asked if they know what the words mean. Before exposing them to the second stimuli, a short explanation is given. In this way, it’s possible to analyze whether results from the first or second stimuli were affected by knowledge.
After having defined what the research questions are and having analyzed the literature available, I proceed with my hypothesis:

H0: *Ad perception is not affected by disclosure nor the use of different words.*
H1: *Disclosure negatively affects consumers’ ad perception*

H2: *If disclosure is positioned at the beginning, or during the ad, consumers will react more negatively than if placed at the end*
H3: *Using "Synthetic Ad" instead of "Deepfake" positively affects the attitude towards the ad.*

Research Methodology and Design

Choosing and editing the media

The research begins with a manual scan of different deepfake videos on YouTube. Specifically for this research, only deepfakes that were applied to an ad were selected. Simultaneously, I reached out by email to the before-mentioned Chris Ume (@vfxchrisume: https://www.instagram.com/vfxchrisume/) CEO and founder of Methaphysics.ai and creator of the already mentioned @deeptomcruise page on TikTok. The reason was the material available on YouTube was not enough hyper-realistic for the study. The videos needed to respect 2 requirements. They should be deepfake ads and they should be hyper-realistic. He reached back and gave explicit consent to use some of his videos. Gillette ad and the Truecaller's ad were the ones chosen for the study.

CU
Chris Ume <chris@metaphysic.ai>
ven 04/02/2022 11:32
A: Andrea Parialò

Hey Andrea,

I've created a few commercials;

You could use this one I did for Truecaller : https://www.youtube.com/watch?v=MiU1WaUNt7I
Or the one I did for Gilette with Deon Sanders (if you google it you will find it)
I also created one where I brought back two legendary managers for the Belgian Football Federation (with consent of the family) at https://www.youtube.com/watch?v=5SYrbjVcuaA&t=36s

Thank you

--

Chris Ume
Co-founder, Metaphysic

M +32484353860
E chris@metaphysic.ai
W https://www.metaphysic.ai/

Figure 38 | email received by Chris Ume

The other 2 videos chosen are a Jaguar spot where YouTube Channel "The Fake Report"[55] substituted Eva Green's face with Angelina Jolie's, and the last one is a L'Oréal spot made by YouTube channel "The Dan In The High Castle"[56] where the main character's face has been substituted with the one of Ana De Armas. All of them look natural and are not as easy to spot. The videos have then been imported into Adobe Premiere Pro where 6 different scenarios for each video were created by positioning a disclaimer message Before, During, or After the ad using either Deepfake or Synthetic Media. The table and pictures below give an example of the disclaimer messages and the different scenarios.

Table 2 | Different Scenarios

	Before ad (B)	During ad (D)	After ad (A)
Deepfake (DF)	B + DF	D + DF	A + DF
Synthetic Media (SM)	B + SM	D + SM	A + SM

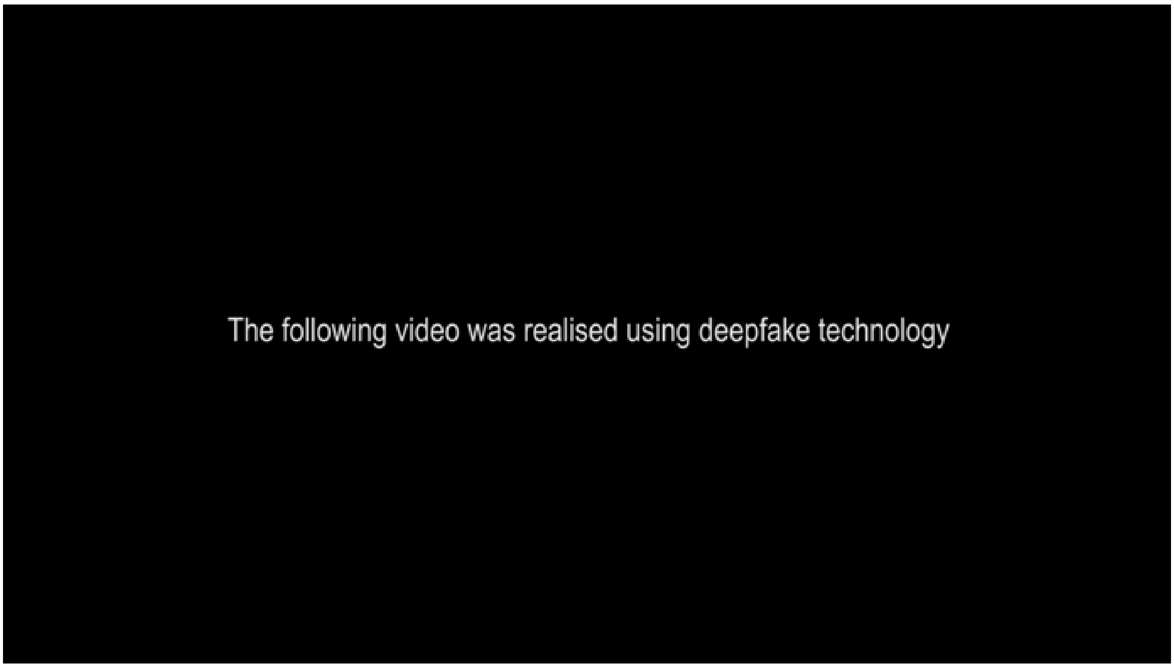


Figure 39 | B + DF

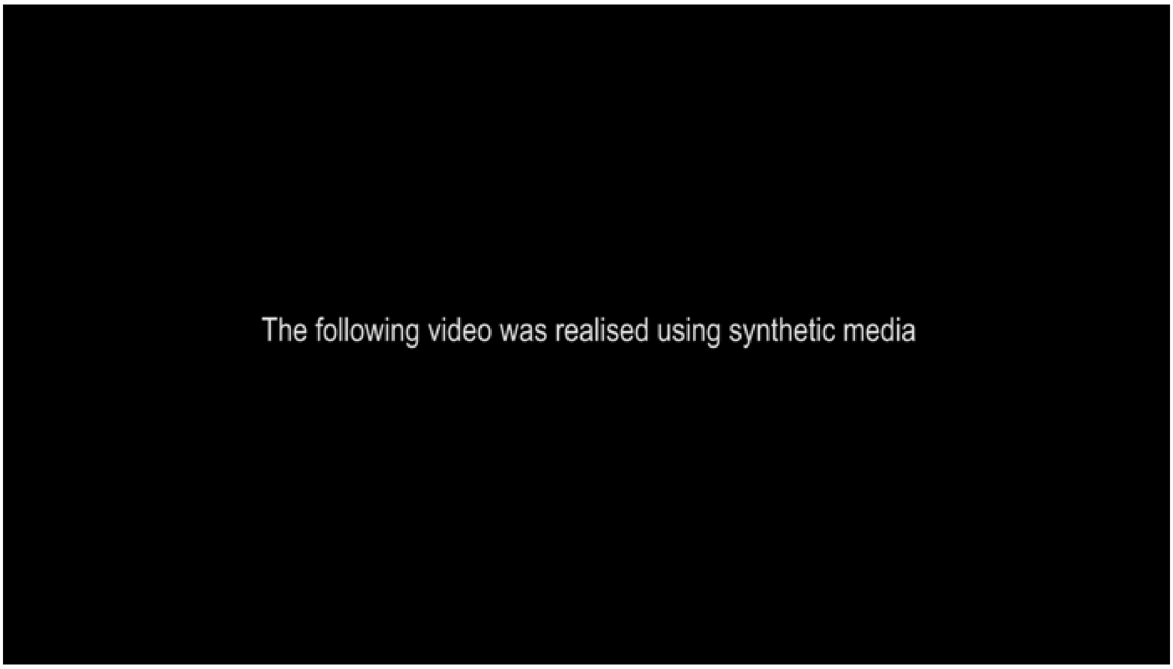


Figure 40 | B + SM

Figure 41 | D + DF

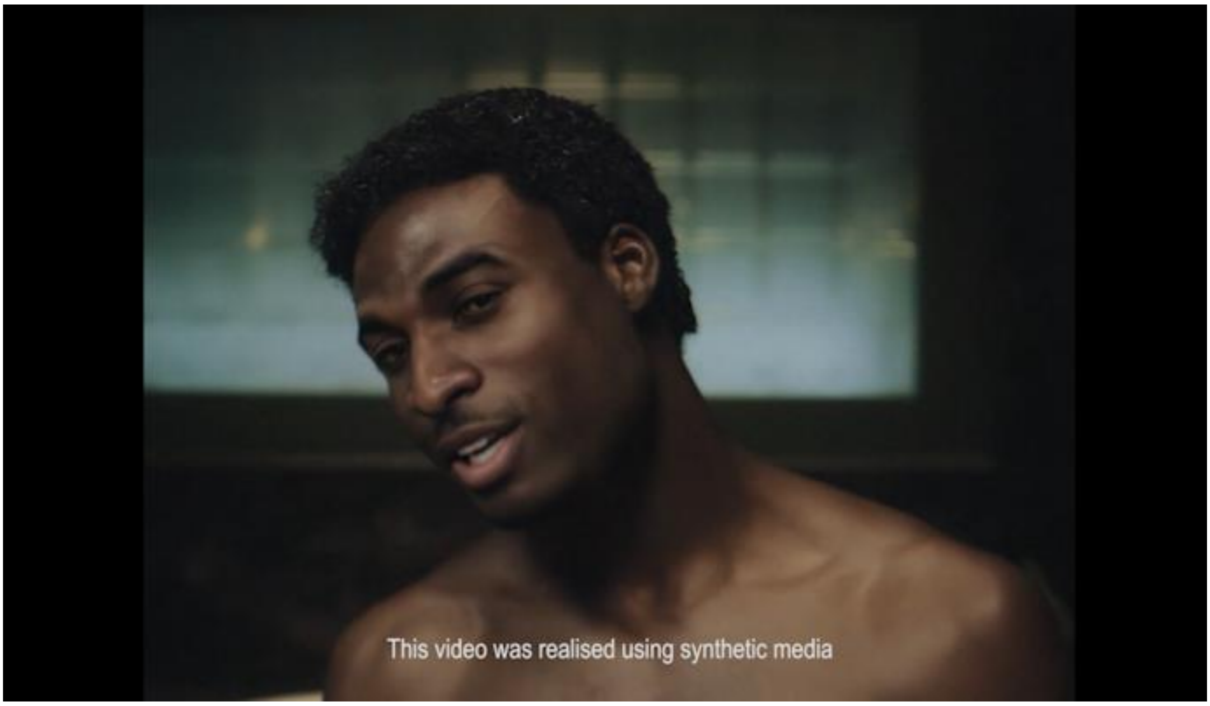


Figure 42 | D + SM

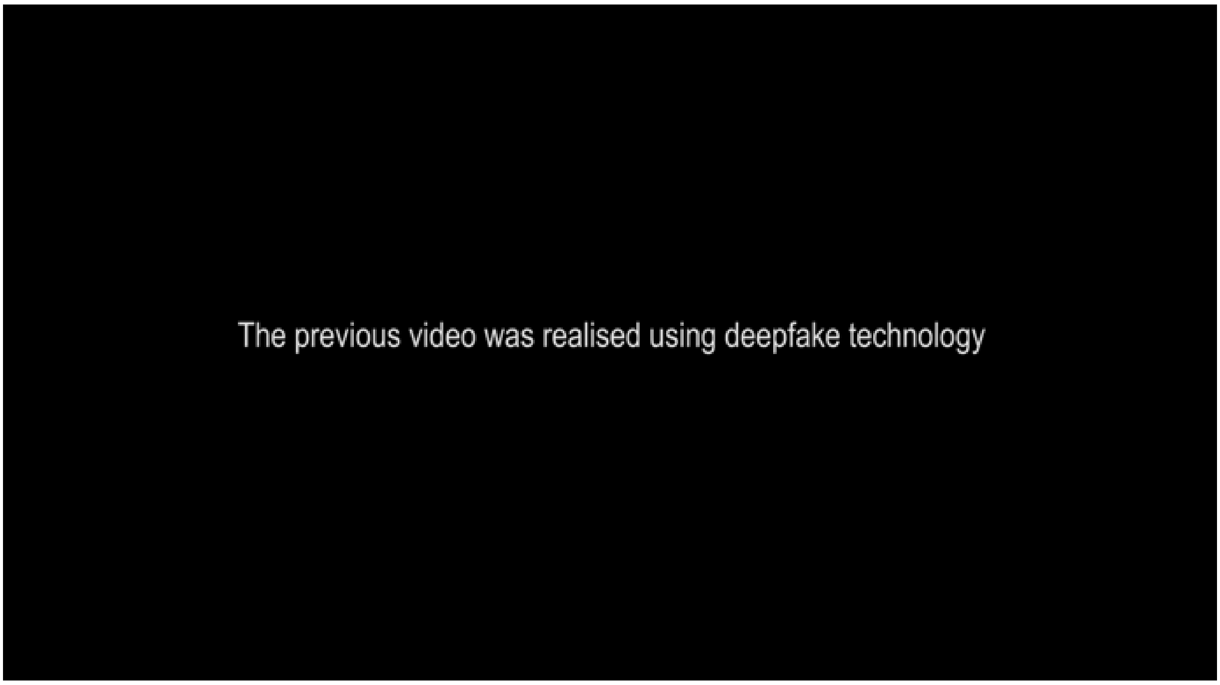


Figure 43 | A + DF

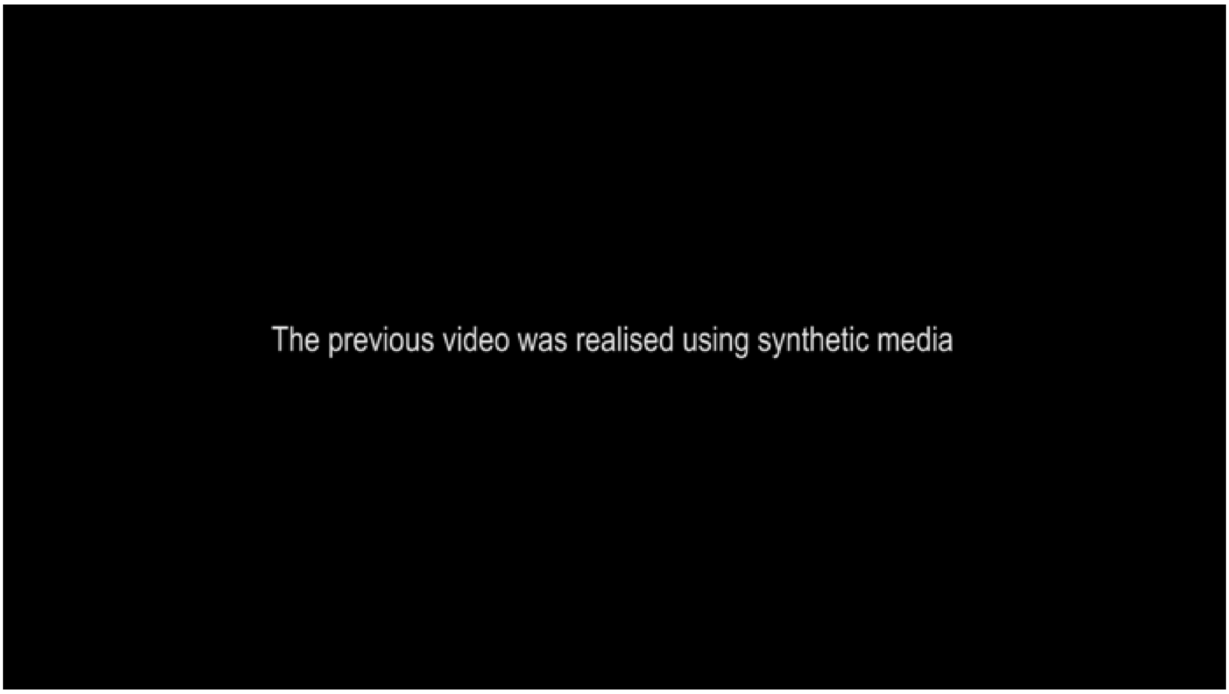


Figure 44 | A + SM

Identification code

To orderly divide and be able to easily recognize the different scenarios. A code has been created for identifying them.

Table 3 | Code Legend

Disclosure:	The disclosure is shown before the ad	B_
	The disclosure is shown during the ad	D_
	The disclosure is shown after the ad	A_
Type*:	Deepfake	DF_
	Synthetic Media	SM_
Brand:	Gillette	GIL
	L'Oréal	LOR
	Jaguar	JAG
	Truecaller	TCC
If one of these variables is missing:	Null variable	N_

The final code is built following a specific order and combining the different single codes for each variable:

[Disclosure] + [Type] + [Brand]

For instance, the code for a scenario where the disclosure is shown during the Truecaller ad, using the term "Synthetic Media" would be:

[D_] + [SM_] + [TCC] = D_SM_TCC.

If any of the variables are missing, "N_" is used as an identifier. So, for example, the code for the original ad by L'Oréal (no disclosure, no deepfake/synthetic media used), would be:

[N_] + [N_] + [LOR] = N_N_LOR.

**Note: If the condition [Disclosure] is NULL (N_), the [Type] part of the code only indicates if the video is a deepfake or not. For example, if a video has no disclosure, so the first part of the code is N_, the code could only continue with either N_ (no manipulation, meaning the video is not a deepfake) or DF_ (manipulation, meaning the video is a deepfake).*

Below are all the possible combinations used in the survey:

Table 4 | Every possible combination created for this study

		L'Oréal	Gillette	Jaguar	Truecaller
Disclosure timing + Type	Before + Deepfake	B_DF_LOR	B_DF_GIL	B_DF_JAG	B_DF_TCC
	Before + Synthetic Media	B_SM_LOR	B_SM_GIL	B_SM_JAG	B_SM_TCC
	During + Deepfake	D_DF_LOR	D_DF_GIL	D_DF_JAG	D_DF_TCC
	During + Synthetic Media	D_SM_LOR	D_SM_GIL	D_SM_JAG	D_SM_TCC
	After + Deepfake	A_DF_LOR	A_DF_GIL	A_DF_JAG	A_DF_TCC
	After + Synthetic Media	A_SM_LOR	A_SM_GIL	A_SM_JAG	A_SM_TCC
	No Disclosure + Manipulated	N_DF_LOR	N_DF_GIL	/	/
	No Disclosure + Original	N_N_LOR	N_N_GIL	/	/

Survey building

After creating the 6 different scenarios, the survey is created on Qualtrics.

Pre-test

After asking what gender the respondents identify in, a pre-test option is shown at the beginning of the main survey.

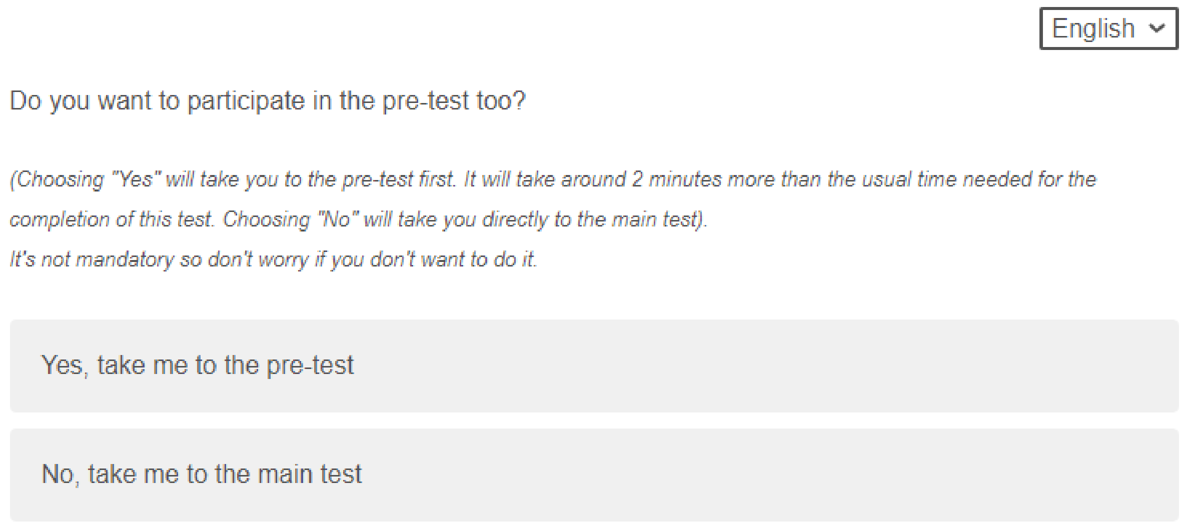


Figure 45 | Screenshot of the pre-test question

If respondents click on "Yes", accepting to participate in the pre-test, they will be redirected to a slightly different journey than the ones who chose "No". To do this, the *Branch Logic* in Qualtrics is used. The branch logic allows you to create an if condition.

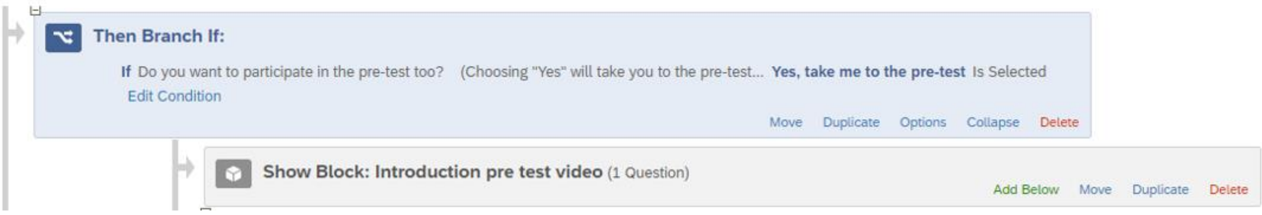


Figure 46 | Branch Logic for pre-test + pre-test introduction block

The purpose of the pre-test is to understand whether the deepfake videos chosen look real enough to trick the general human eye. The introduction block disclaims that a video would be played afterward, and informs the respondent has a 50% chance of watching a real video, and a 50% chance to watch a fake video.

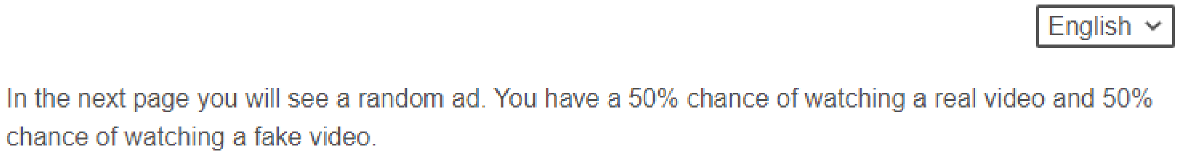


Figure 47 | Pre-test introduction block

The videos chosen for the pre-test are:

- N_N_LOR
- N_N_GIL
- N_DF_LOR
- N_DF_GIL

To avoid any contamination of the subsequent main test, the gender variable functions as a divider for making sure a respondent does not watch the same video twice. If you are male, you will watch the L'Oréal ad (either deepfake or actual one) in pre-testing, and the Gillette one during the main test. If you are female, the opposite. If you are non-binary or prefer not to say, you will be randomly assigned a video between the 2 types of ads. In this way, we make sure to create no bias towards the ads that will be shown later in the survey. The figure below gives an example of the flow for a male.

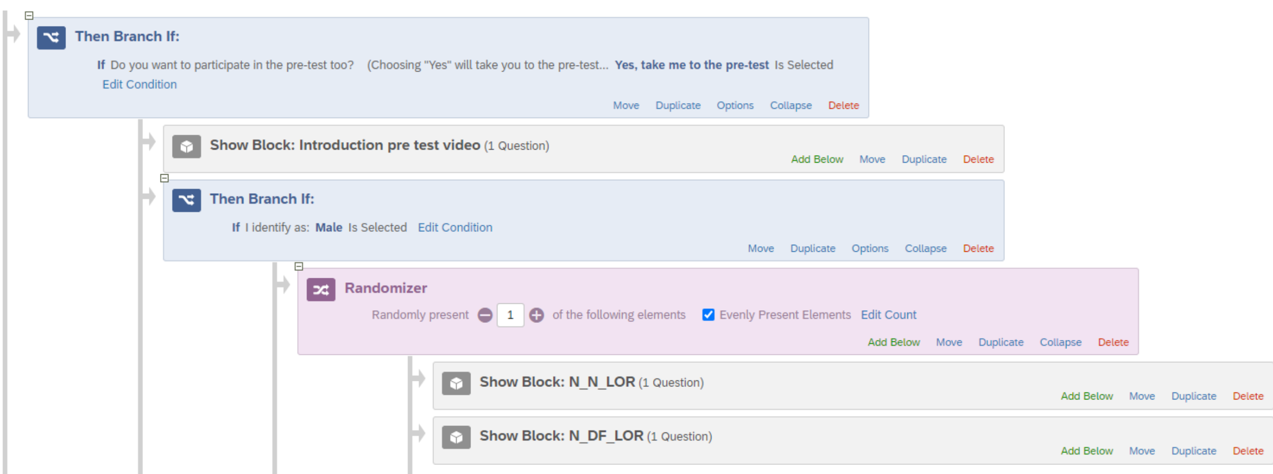


Figure 48 | Male flow for pre-testing

A randomizer is used to expose the respondent to either one of the 2 stimuli (not deepfake vs deepfake). In the case shown in the previous figure, as explained before, N_N_LOR indicates the video has no disclosure (N_), is not a deepfake (N_), and is from L'Oreal (LOR). Conversely, N_DF_LOR indicates the video has no disclosure (N_), is a deepfake (DF_), and is from L'Oreal (LOR).
The videos chosen for this purpose are Gillette's ad "Ready or Not" which will be tested against the deepfake ad created by Chris Ume for Gillette's "Deion's Draft Night" ad (deepfake).

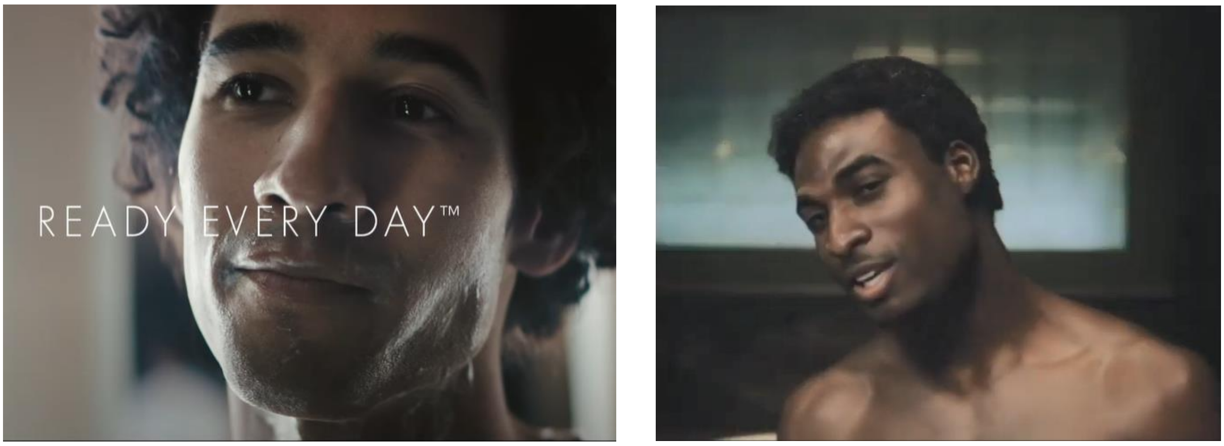


Figure 49 | Gillette's Ready or Not ad (N_N_GIL - left) vs Gillette's Deion's Draft Night (N_DF_GIL - right)

On the other side, male participants will be presented with an identical scenario. The only difference is the type of ad. In this case, one is the original 1989 spot by L'Oréal starring Cindy Crawford while the other is a deepfake of the same ad with Ana De Armas.

Figure 50 | Original spot (N_N_LOR - left) vs Deepfake (N_DF_LOR - right)

To evenly expose participants to every scenario, the *Randomizer* function is used to either display the deepfake version of an ad, or its equivalent real one. After being exposed to the randomly assigned 30-seconds spot, respondents are asked to answer a specific and clear question shown in figure 51.

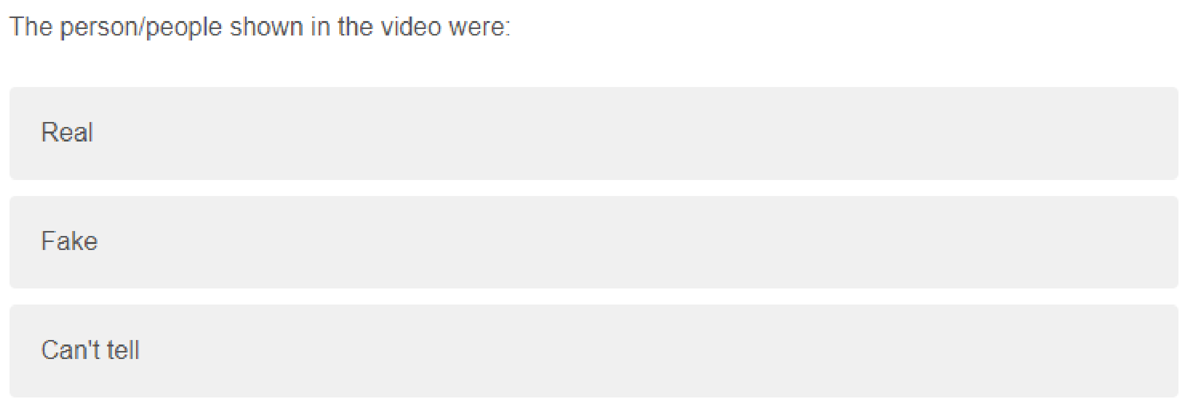


Figure 51 | Question after watching the video in the pre-test

After measuring this, a message saying “you will now start the main test” will be shown before redirecting them to the main test.

Main Test

The main test begins with a message saying they will watch an ad on the next page.
The survey flow from now on follows the same concept as depicted before. The variable gender is used to control what scenario respondents are going to be exposed to. In this case, and contrariwise to the pre-test, if “Male” is chosen they will be shown the Gillette ad. If “Female” is chosen they will be shown the L’Oréal ad. If “Non-Binary/Third Gender” or “Prefer not to say” is chosen they would be randomly assigned to one of the two videos.
The reason behind this choice is the target for the two ads are different. Gillette’s target is men, L’Oréal’s target is women.
An ad without any disclaimer for both Gillette and L’Oréal has been added in the randomization to be used as a litmus test for comparison between an ad with a disclaimer and without.

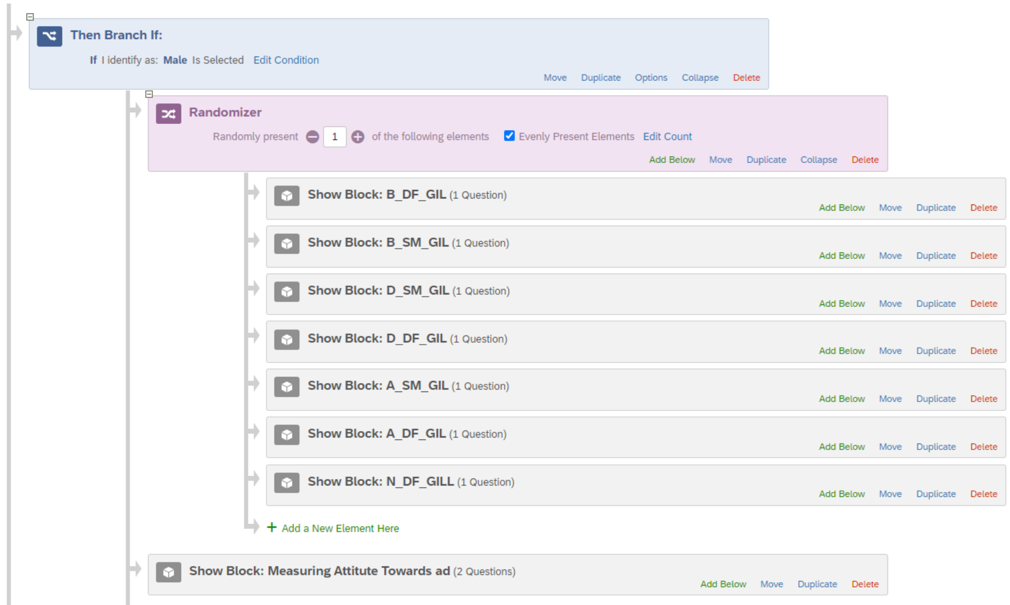


Figure 52 | Survey flow for male

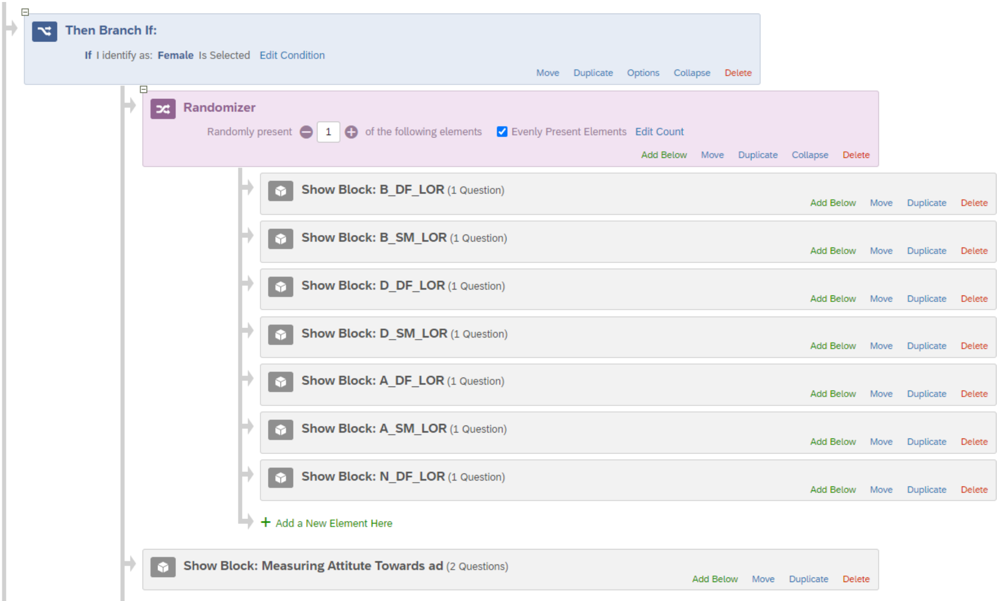


Figure 53 | Survey flow for female

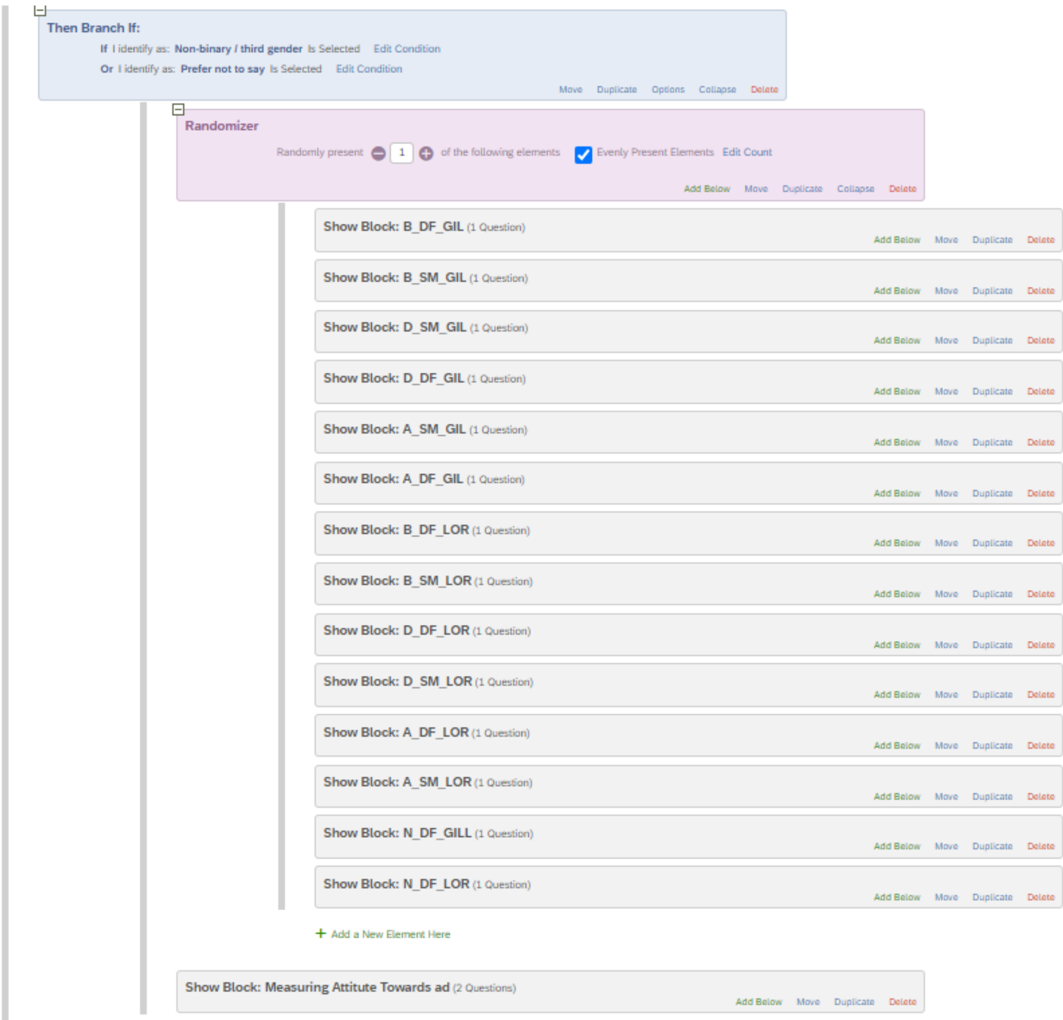
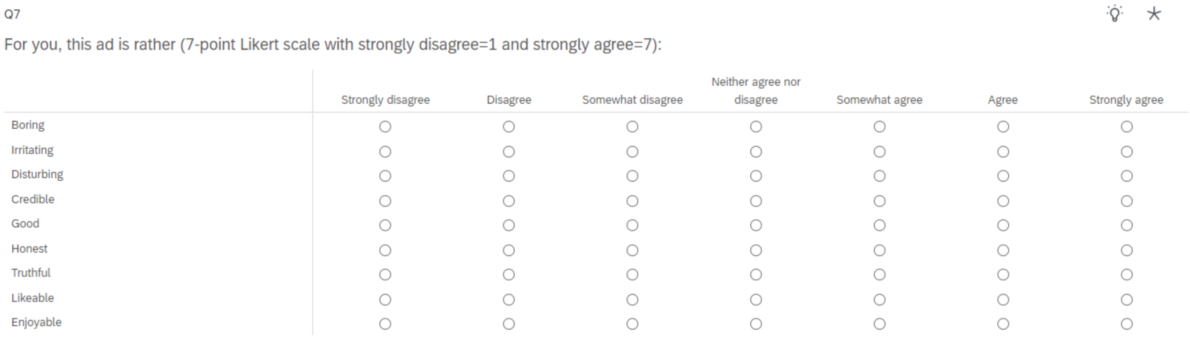


Figure 54 | Survey Flow for NonBinary/ThirdGender

The ADTRUST scale was chosen to test the attitudes towards the ad (Soh, Reid, and King 2007) and items thought to be a better fit for the research were selected. The scale is a 7-point Likert scale (with Strongly Disagree = 1 and Strongly Agree = 7) that measures how much the respondent agrees to some adjectives describing the ads that are the items of the questionnaire.
The items are: Boring, Irritating, Disturbing, Credible, Good, Honest, Truthful, Likeable, Enjoyable.

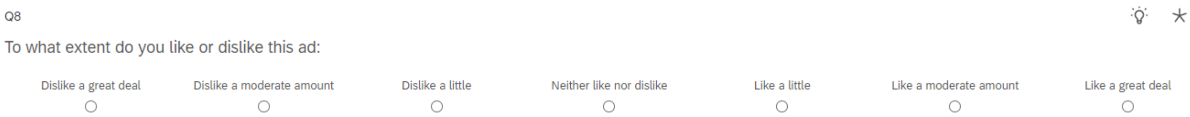

Q7

For you, this ad is rather (7-point Likert scale with strongly disagree=1 and strongly agree=7):

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
Boring	○	○	○	○	○	○	○
Irritating	○	○	○	○	○	○	○
Disturbing	○	○	○	○	○	○	○
Credible	○	○	○	○	○	○	○
Good	○	○	○	○	○	○	○
Honest	○	○	○	○	○	○	○
Truthful	○	○	○	○	○	○	○
Likeable	○	○	○	○	○	○	○
Enjoyable	○	○	○	○	○	○	○

Figure 55 | ADTRUST

Right after ADTRUST, another question measuring to what extent they generally liked the ad is asked. The question asks “To what extent do you like or dislike this ad: (7-point Likert scale with like = 7; not like = 1)” (MacKenzie, Lutz, and Belch 1986).


Q8

To what extent do you like or dislike this ad:

Dislike a great deal	Dislike a moderate amount	Dislike a little	Neither like nor dislike	Like a little	Like a moderate amount	Like a great deal
○	○	○	○	○	○	○

Figure 56 | Last question

Once finished, the respondents will be randomly split into 2 groups. The group “Deepfake” and the group “Synthetic Media”. Group Deepfake, for example, will be asked first if they know what a Deepfake is, giving them only two options: “Yes” and “No”. After answering the question, a message with a short Deepfake definition, and an easy question about the definition itself are shown. This step has a double function. It will both give the information needed to people who might not know what the term meant, and function as a check for cheaters/filter out invalid responses. The figure below better explains the concept.

Deepfakes are a technique used to create media in which a person in an existing image or video is replaced with someone else's likeness.

Deepfakes are a technique used to create:

paper planes

media in which a person in an existing image or video is replaced with someone else's likeness.

iron boats

Figure 57 | Information given + cheater checker

Once answered, the survey shows group Deepfake another ad. This time, chosen from the DF_TCC and DF_JAG ads. Disclosure timing is the only randomized variable. After that, the same questions used before for measuring consumers' attitudes are asked. The same thing happens to group "Synthetic Media", the only difference is the variable [Terminology], which in this case will be SM_.

After answering this last question, the survey is considered completed and a thank you message is displayed.

Data Analysis

The R script is available in the appendix at the end of the document. Before starting to analyze the raw data on R, a preliminary clean-up and sorting of the dataset were done on Microsoft Excel[57]. The survey received 377 responses. The first thing checked was the "cheater filter" question. The observations that didn't pass the test were deleted from the dataset. The next thing was setting up a threshold for a minimum time required to properly do the survey. I did the survey several times to have an average time of completion that was roughly 300 seconds if taking the Pre-Test, and 150 seconds if only doing the Main Test. I then proceeded to delete all the observations that didn't pass the cut.
Still on Excel, columns for the different scenarios were renamed using the code explained before for identifying them.
The updated dataset was then uploaded into RStudio. After some final data cleaning and adjustments on R and having divided the dataset into Pre-Test and Main-Test, a first demographic overview is conducted.

World map

A World Map is plotted to have an overview of where the respondents come from.

Respondents Map Overview

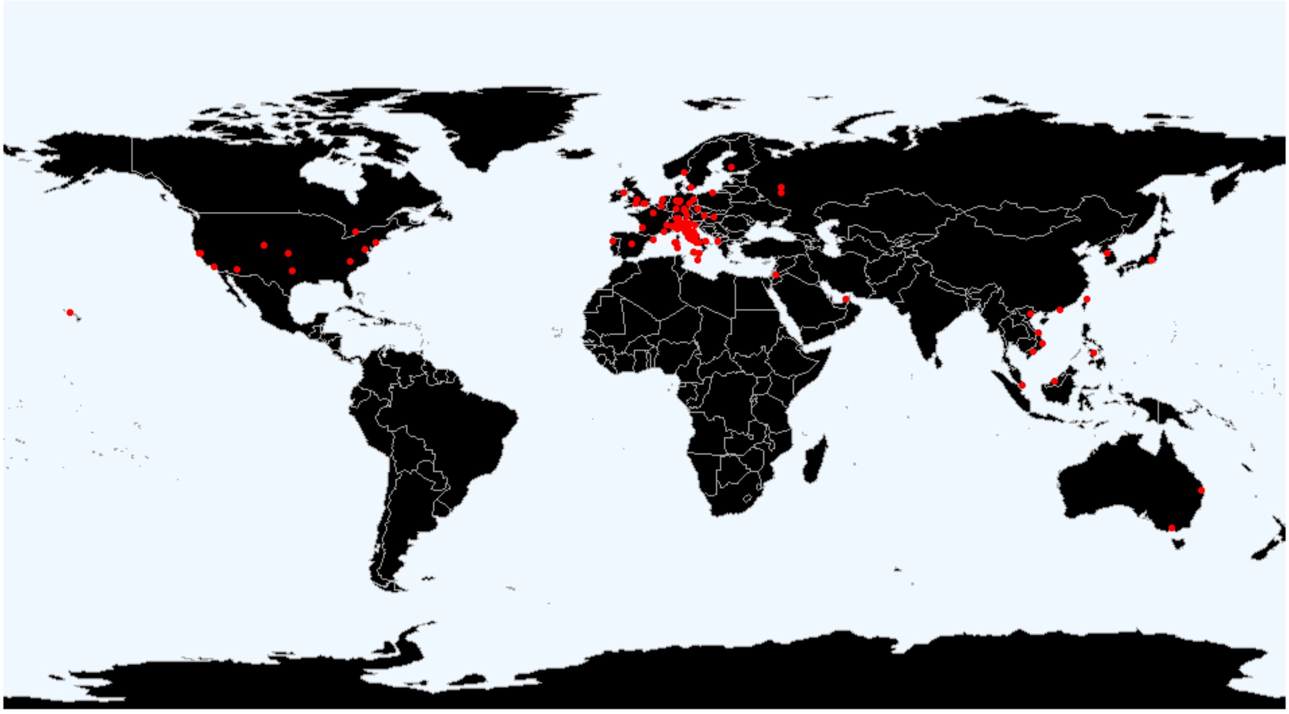


Figure 58 | Plotted map of respondents who participated in the study

Pre-Test

155 out of the total 300 participants involved in the study, participated in the pre-test. The main of the pre-test was to test whether the videos chosen were enough real to trick the human eye.

Pre-test demographics

```
> summary(PT_Demographics)
                          EMPLOYMENT                  EDUCATION        AGE                                 Gender
 Employed full time            :77   2ndL Master         : 7   18 - 24:56   Female                     :91
 Employed part time            :16   Bachelor            :59   25 - 34:92   Male                       :57
 Retired                       : 1   HighSchool          :27   35 - 44: 3   Non-binary / third gender: 4
 Student                       :50   Postgraduate        :56   45 - 54: 3   Prefer not to say          : 3
 Unemployed looking for work   : 9   Professional degree: 6   55 - 64: 1
 Unemployed not looking for work: 2
```

Figure 59 | summary() result for pre test demographics

PT_Demographics' summary gives an overview of the composition of the pre-test sample composition.

Pre-test results

Results from the Pre-Test show a high percentage of failing to spot the deepfake.

```
, , BRAND = GIL

                    N_DF
PRE_TEST.REAL_or_FAKE DF  N
          Can't tell 12  5
          Fake       13  7
          Real       22 35

, , BRAND = LOR

                    N_DF
PRE_TEST.REAL_or_FAKE DF  N
          Can't tell  3  2
          Fake        3  0
          Real       26 27
```

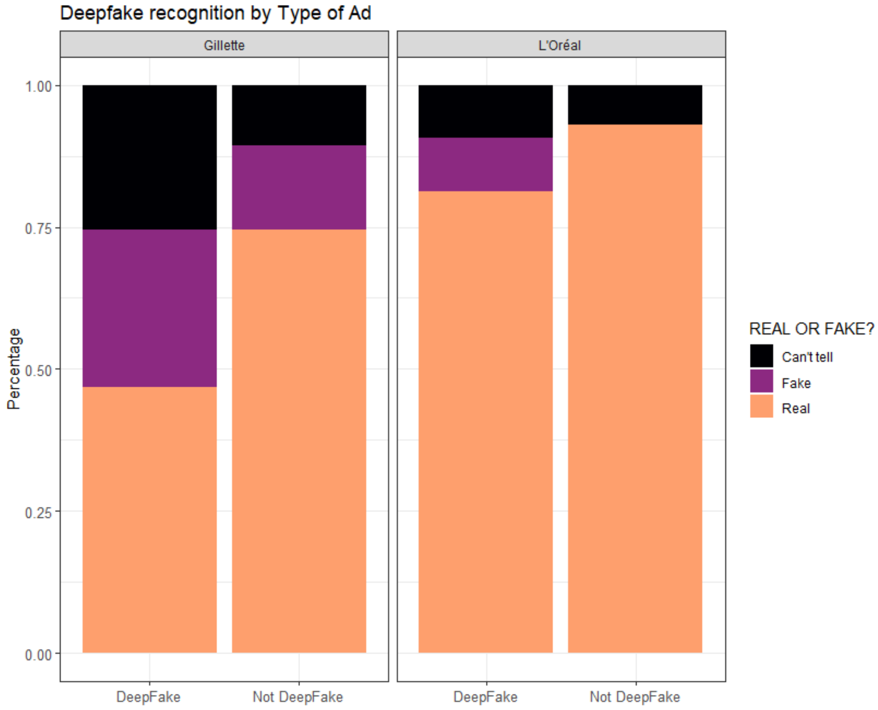


Figure 60 | Pre-Test results

For Brand Gillette, the deepfake video passed the human eye check and got marked as real 46.8% of the time and only 27.7% of the time was correctly spotted as fake, while 25.5% of the time respondents were not able to accurately say if it was real or not. The "not deepfake" Gillette ad got wrongly marked as fake 14.9% of the time. What performed incredibly well, though, was the L'Oréal deepfake, which managed to trick the human eye 81.2% of the time. This is much less than just flipping a coin. Only 9.4% of the time the video was correctly spotted as fake.
The results are surprisingly positive, and the pre-test was successful. The videos are realistic enough for conducting the study.

Main Test

Demographics overview

```
> summary(MT_Demographics)
                        EMPLOYMENT                  EDUCATION        AGE                              Gender
 Employed full time              :132   2ndL Master          : 14   18 - 24 :117   Female                    :184
 Employed part time              : 37   Bachelor             :117   25 - 34 :168   Male                      :105
 Retired                         :  1   HighSchool           : 45   35 - 44 :  6   Non-binary / third gender:  5
 Student                         :107   No HighSchool        :  1   45 - 54 :  7   Prefer not to say         :  6
 Unemployed looking for work     : 17   PhD                  :  4   55 - 64 :  1
 Unemployed not looking for work:  6   Postgraduate         :110   Under 18:  1
                                        Professional degree:  9
```

Figure 61 | summary() result for Main Test demographics

The sample composition for the main test differs a lot but has some more prominent variables. The variable gender included – Female (61.3%), Male (35%), Prefer Not to Say (2%), Nonbinary / Third gender (1.7%). Most of the respondents' ages range between 25-34 (56%) and 18-24 (39%). Education-wise the most numerous have a bachelor's degree (39%), a Postgraduate Degree, or equivalent (36.7%). And finally, they are mostly Employed full time (44%) or are students (35.7%).

PCA (Principal Component Analysis) – Loading Plot

A PCA is conducted to find the first two PCs that together explain the 64.8% of the variance. We can have a first overview of our results. We see how the first three columns “AD_BORING_1”, “AD_IRRITATING_1”, and “AD_DISTURBING_1” are similar in measuring a negative feeling towards the ad like we expected. On the contrary, we notice how all of the others are negatively correlated to the previous ones. At the same time, we notice “AD_HONEST_1”, “AD_TRUTHFUL_1”, “AD_CREDIBLE_1” are all correctly going in the same positive direction but slightly differently than “AD_GOOD_1”, “LIKE_1”, “AD_ENJOYABLE_1”, and “AD_LIKEABLE_1” which again, give us an insight that the survey is consistent.

Figure 62 | PCA for exposition 1

Similarly, PCA analysis for exposition 2 confirms what emerged from the first PCA. The directions of the arrows are very close to the ones for exposition 1.

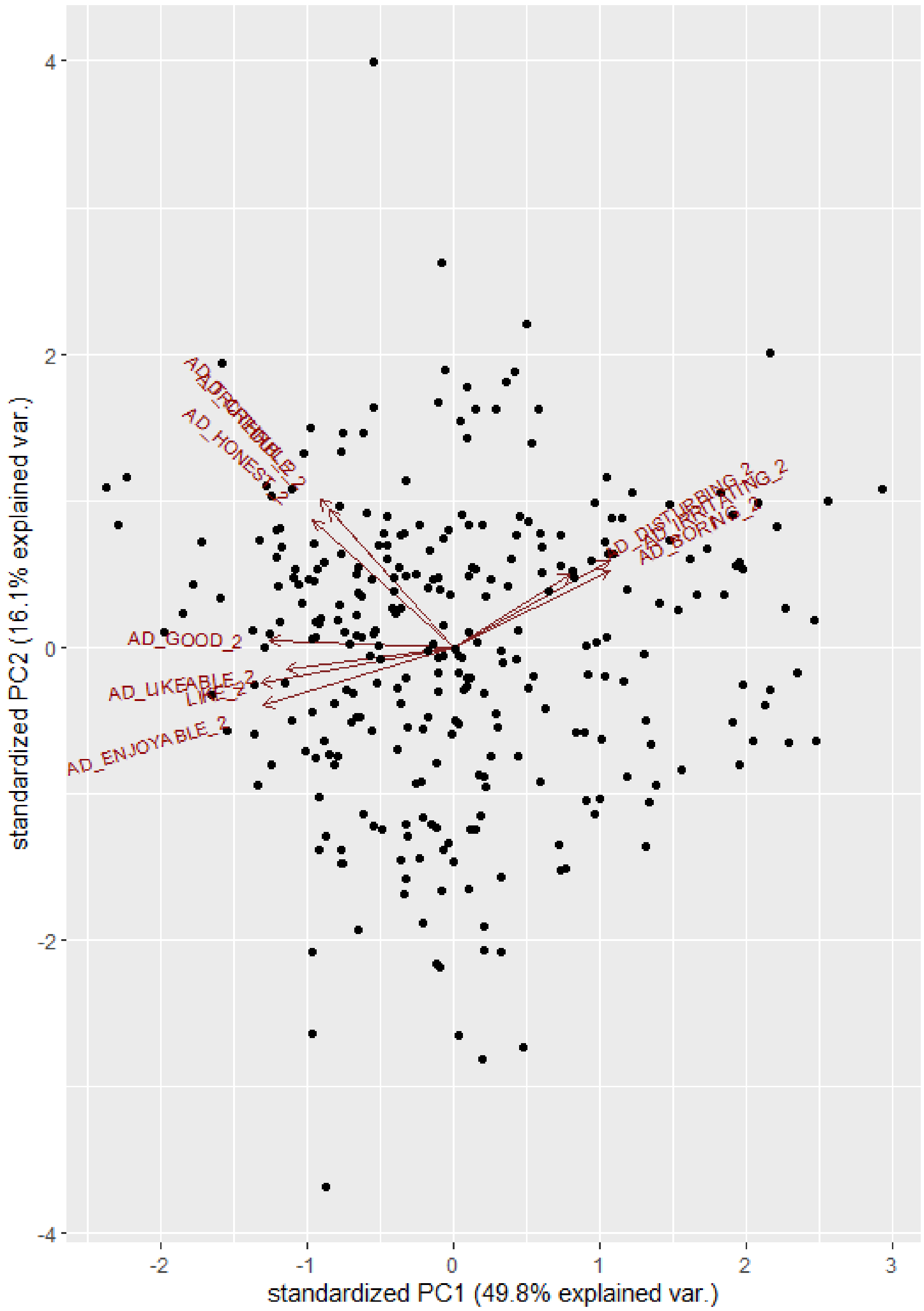


Figure 63 | PCA for exposition 2

Cronbach's alpha

After the PCA is done, Cronbach's alpha is checked for both expositions to see if the scale used is reliable and if some items need to be dropped. To do so, the first 3 items are inverted being negatively correlated to the rest.
The results for exposition 1 are positive with an alpha of .9 and no further improvements if an item is dropped. Being .9 > .7 we accept the scale as reliable (Kline 2000).

```
Reliability analysis
Call: psych::alpha(x = reliability_1)

  raw_alpha std.alpha G6(smc) average_r S/N    ase mean  sd median_r
       0.9       0.9    0.91      0.47 8.7 0.0091  3.8 1.1     0.47

 lower alpha upper     95% confidence boundaries
0.88 0.9 0.91

 Reliability if an item is dropped:
                raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
AD_BORING_1          0.89      0.89    0.90      0.48 8.3   0.0096 0.017  0.46
AD_IRRITATING_1      0.88      0.89    0.90      0.47 7.9   0.0101 0.021  0.45
AD_DISTURBING_1      0.90      0.90    0.91      0.50 9.1   0.0089 0.015  0.49
AD_CREDIBLE_1        0.89      0.89    0.90      0.47 7.9   0.0100 0.020  0.47
AD_GOOD_1            0.88      0.88    0.89      0.45 7.3   0.0107 0.018  0.44
AD_HONEST_1          0.88      0.89    0.89      0.46 7.8   0.0101 0.018  0.47
AD_TRUTHFUL_1        0.89      0.89    0.90      0.48 8.3   0.0096 0.017  0.48
AD_LIKEABLE_1        0.88      0.89    0.90      0.46 7.7   0.0103 0.018  0.46
AD_ENJOYABLE_1       0.88      0.88    0.89      0.45 7.4   0.0107 0.017  0.46
LIKE_1               0.88      0.88    0.89      0.44 7.2   0.0108 0.016  0.45
```

Figure 64 | Cronbach's alpha for exposition 1

Similarly, exposition 2 confirms again what we found from analyzing exposition 1. Alpha .88 > .7, we accept the scale as reliable.

```
Reliability analysis
Call: psych::alpha(x = reliability_2, check.keys = TRUE)

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean sd median_r
      0.88      0.88    0.91      0.43 7.6 0.011  4.6  1     0.41

 lower alpha upper     95% confidence boundaries
0.86 0.88 0.9

 Reliability if an item is dropped:
                raw_alpha std.alpha G6(smc) average_r S/N alpha se var.r med.r
AD_BORING_2          0.87      0.87    0.90      0.43 6.9    0.012 0.028  0.41
AD_IRRITATING_2      0.87      0.87    0.90      0.43 6.8    0.012 0.030  0.39
AD_DISTURBING_2      0.88      0.88    0.91      0.46 7.7    0.011 0.028  0.43
AD_CREDIBLE_2        0.88      0.88    0.90      0.45 7.4    0.011 0.027  0.42
AD_GOOD_2            0.86      0.86    0.89      0.41 6.2    0.013 0.030  0.38
AD_HONEST_2          0.87      0.88    0.90      0.44 7.0    0.011 0.030  0.42
AD_TRUTHFUL_2        0.87      0.88    0.89      0.44 7.2    0.011 0.027  0.43
AD_LIKEABLE_2        0.86      0.87    0.90      0.42 6.6    0.012 0.031  0.41
AD_ENJOYABLE_2       0.85      0.86    0.88      0.41 6.1    0.013 0.026  0.39
LIKE_2               0.85      0.86    0.89      0.40 6.1    0.013 0.026  0.40
```

Figure 65 | Cronbach's alpha for exposition 2

Model fitting

Variables AD_TRUST_1 and AD_TRUST_2 are created to represent the total score of the scale used to measure the attitude towards the ad. Each score is calculated by averaging the items' values for each observation and subsequently converting them into a 1 to 100 scale.
The model we began with, translating the research question was:

AD_TRUST = DISCLOSURE * TYPE

To have a better overview and understanding, different combinations of models are used to find the best match.

Exposition 1:

Testing different types of ANOVA (exposition 1)

The models analyzed are:

One-way ANOVA: AD_TRUST_1 = DISCLOSURE_1

Two-way ANOVA: AD_TRUST_1 = DISCLOSURE_1 * TYPE_1

Two-way ANOVA + blocking variable: AD_TRUST_1 = DISCLOSURE_1 * TYPE_1 + BRAND_1

Three-way ANOVA: AD_TRUST_1 = DISCLOSURE_1 * TYPE_1 * value

Three-way ANOVA + blocking variable: AD_TRUST_1 = DISCLOSURE_1 * TYPE_1 * value + BRAND_1

Four-way ANOVA: AD_TRUST_1 = DISCLOSURE_1 * TYPE_1 * value * BRAND_1

"value" = previous knowledge of deepfake/synthetic media

The aictab() function is then used for comparing the different models and finding the best-fit model doing an Akaike Information Criterion (AIC) test.

```
Model selection based on AICc:

                 K    AICc Delta_AICc AICcWt Cum.Wt       LL
2W_BLOCK_ANOVA   9 2449.02       0.00   0.96   0.96 -1215.20
3W_BLOCK_ANOVA  16 2455.56       6.55   0.04   1.00 -1210.82
4W_ANOVA        29 2468.26      19.24   0.00   1.00 -1201.91
1W_ANOVA         5 2491.97      42.96   0.00   1.00 -1240.88
2W_ANOVA         8 2497.47      48.45   0.00   1.00 -1240.49
3W_ANOVA        15 2502.37      53.35   0.00   1.00 -1235.34
```

Figure 66 | AIC test, exposition 1

The result lists the models in order of best fit. In this case, the two-way ANOVA + blocking variable is the best-fit one of the 6 tested. The lowest AICc is the best. In this case, the lowest AICc score is 2449.02. The variable AICcWt shows the proportion of the total amount of predictive power provided by the full set of models analyzed. 96% of the total explanation can be found in the full set of models. (Bevans 2020). For these reasons, the model chosen is the 2W_BLOCK_ANOVA. The result of the chosen model is shown below:

```
                   Df Sum Sq Mean Sq F value    Pr(>F)    
DISCLOSURE_1        3   2504     835   4.207    0.0062 ** 
TYPE_1              1     17      17   0.087    0.7677    
BRAND_1             1  10601   10601  53.424 2.59e-12 ***
DISCLOSURE_1:TYPE_1  2    204     102   0.514    0.5984    
Residuals         292  57944     198                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 67 | 2W_BLOCK_ANOVA, exposition 1

From this result, we notice that BRAND_1 is the more responsible for changing the value of AD_TRUST_1 with a p-value < .001. The second most important variable is DISCLOSURE_1, with a p-value < .01. Given these results, we can reject the NULL hypothesis.

Testing the model for homoscedasticity (exposition 1)

Checking if the model fits homoscedasticity by looking at the diagnostic plots.

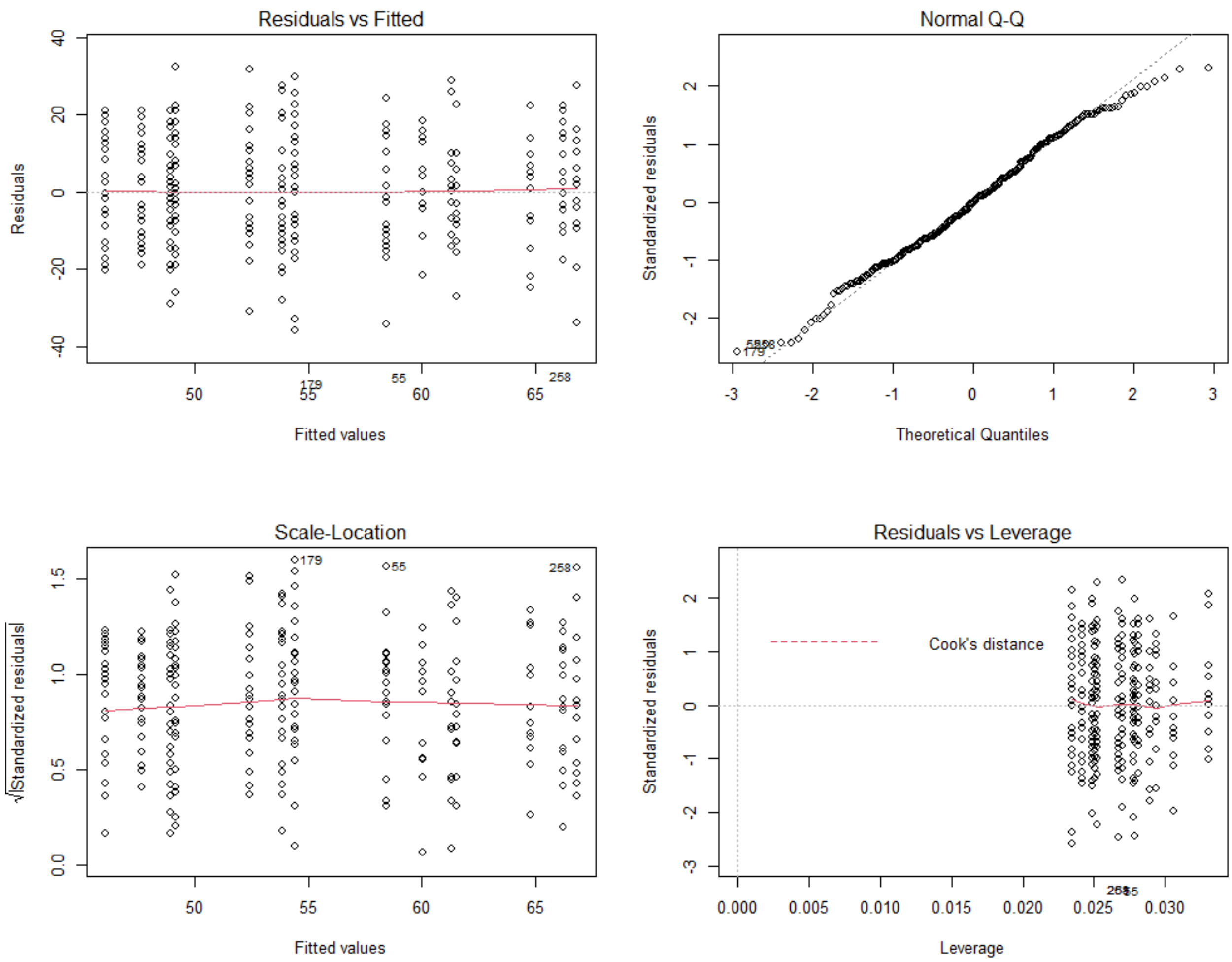


Figure 68 | Diagnostic plot model, exposition 1

We want to put focus on the red centered lines representing the mean of the residuals. If, like in this case, they are horizontal and centered to zero (or one in the scale-location plot), it means we don't have large outliers that will bias the model. The model fits the assumption of homoscedasticity.

Post-hoc test (exposition 1)

To further investigate which groups, differ the most in the variables we run a Turkey's Honestly Significant Difference (Turkey's HSD) post-hoc test.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = ADTRUST_1 ~ DISCLOSURE_1 * TYPE_1 + BRAND_1, data = MAIN_TEST)

$DISCLOSURE_1
          diff         lwr        upr     p adj
B-A  4.9753695  -0.6094623 10.5602013 0.0999281
D-A -0.9183673  -6.5516800  4.7149453 0.9748221
N-A  5.8695652  -0.8209753 12.5601057 0.1082628
D-B -5.8937368 -11.4615300 -0.3259436 0.0333069
N-B  0.8941958  -5.7412727  7.5296642 0.9854764
N-D  6.7879326   0.1116082 13.4642569 0.0446320

$TYPE_1
           diff       lwr      upr     p adj
SM-DF 0.4542797 -2.792475 3.701034 0.7832223

$BRAND_1
             diff       lwr      upr p adj
LOR-GIL -12.32573 -15.64736 -9.00411     0

$`DISCLOSURE_1:TYPE_1`
                 diff        lwr       upr     p adj
B:DF-A:DF   3.4752703  -5.972299 12.922840 0.9513885
D:DF-A:DF  -3.0103263 -12.308242  6.287589 0.9758190
N:DF-A:DF   4.8803418  -4.417573 14.178257 0.7485732
A:SM-A:DF  -1.9094312 -11.357001  7.538138 0.9986357
B:SM-A:DF   4.4854111  -4.910277 13.881099 0.8291638
D:SM-A:DF  -0.5727003 -10.315208  9.169808 0.9999997
N:SM-A:DF          NA         NA        NA        NA
D:DF-B:DF  -6.4855966 -15.608375  2.637182 0.3727845
N:DF-B:DF   1.4050715  -7.717707 10.527850 0.9997712
A:SM-B:DF  -5.3847015 -14.659960  3.890557 0.6392667
B:SM-B:DF   1.0101408  -8.212267 10.232549 0.9999772
D:SM-B:DF  -4.0479706 -13.623477  5.527536 0.9018325
N:SM-B:DF          NA         NA        NA        NA
N:DF-D:DF   7.8906681  -1.077038 16.858374 0.1309140
A:SM-D:DF   1.1008951  -8.021883 10.223674 0.9999558
B:SM-D:DF   7.4957373  -1.573302 16.564776 0.1897499
D:SM-D:DF   2.4376260  -6.990257 11.865509 0.9935739
N:SM-D:DF          NA         NA        NA        NA
A:SM-N:DF  -6.7897730 -15.912551  2.333005 0.3125635
B:SM-N:DF  -0.3949308  -9.463970  8.674108 1.0000000
D:SM-N:DF  -5.4530421 -14.880925  3.974841 0.6436840
N:SM-N:DF          NA         NA        NA        NA
B:SM-A:SM   6.3948422  -2.827566 15.617250 0.4064252
D:SM-A:SM   1.3367309  -8.238776 10.912237 0.9998816
N:SM-A:SM          NA         NA        NA        NA
D:SM-B:SM  -5.0581114 -14.582433  4.466210 0.7370791
N:SM-B:SM          NA         NA        NA        NA
N:SM-D:SM          NA         NA        NA        NA
```

Figure 69 | Turkey's HSD test, exposition 1

From this test, we compare the different scenarios one with another to see which differences are statistically significant. The only significant differences emerging from this test are between "During" (D_) and "Before" (B_) (p-value < .05), and between "During"(D_) and "None"(N_) (p-value < .05) for DISCLOSURE_1.
There are no significant differences between TYPE_1 itself nor between the interaction of TYPE_1 with DISCLOSURE_1, suggesting there is no moderating effect. We then plot the result:

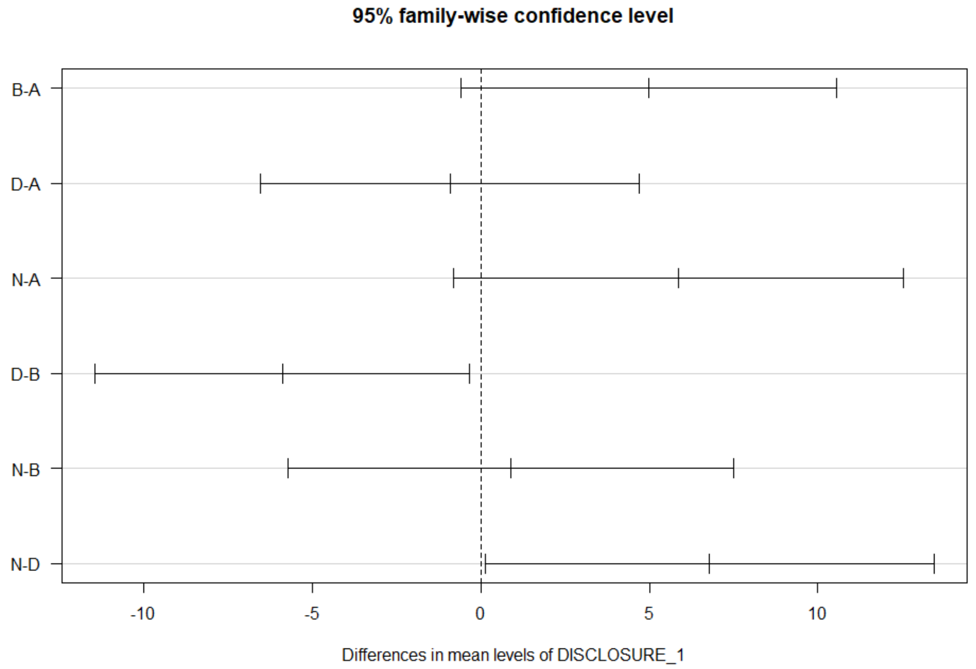


Figure 70 | the plotted result

An interaction plot is then plotted to have a clearer overview of how the variables interact with each other:

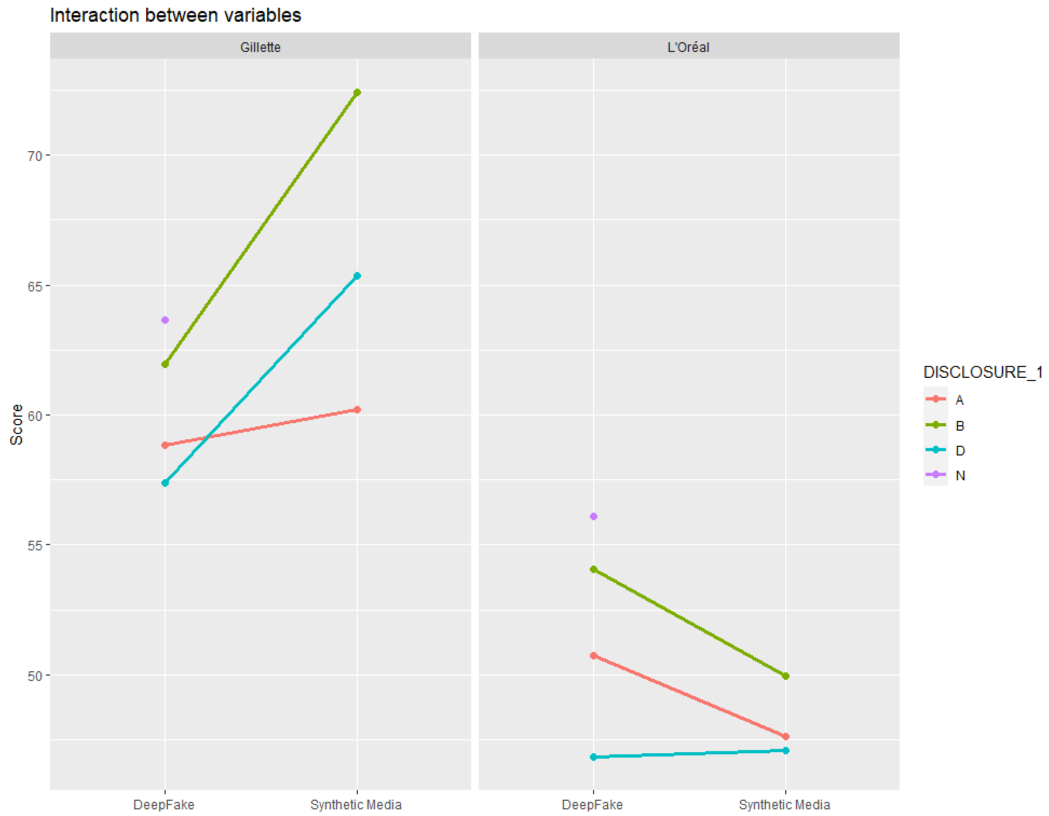


Figure 71 | Exposition_1 interaction plot

From this plot, we can immediately notice how the total score (AD_TRUST_1) is firstly influenced by the type of ad (BRAND_1), which understandably plays a big role in the level of positive attitude towards the ad. Many are the variables involved that affect how a person perceives it. In line with what was expected, no disclosure results in a higher score, but conversely to what was hypothesized, if the disclosure is placed before (B), the score seems to be higher compared to the other three disclosure timings. Lastly, the terminology used (TYPE_1) doesn't seem to affect much the outcome, which seems to be unpredictable or in some way influenced by the type of brand, which seems to be illogical.

Overview plot for exposition 1

After validating the model, an overview plot is created to summarize the differences.

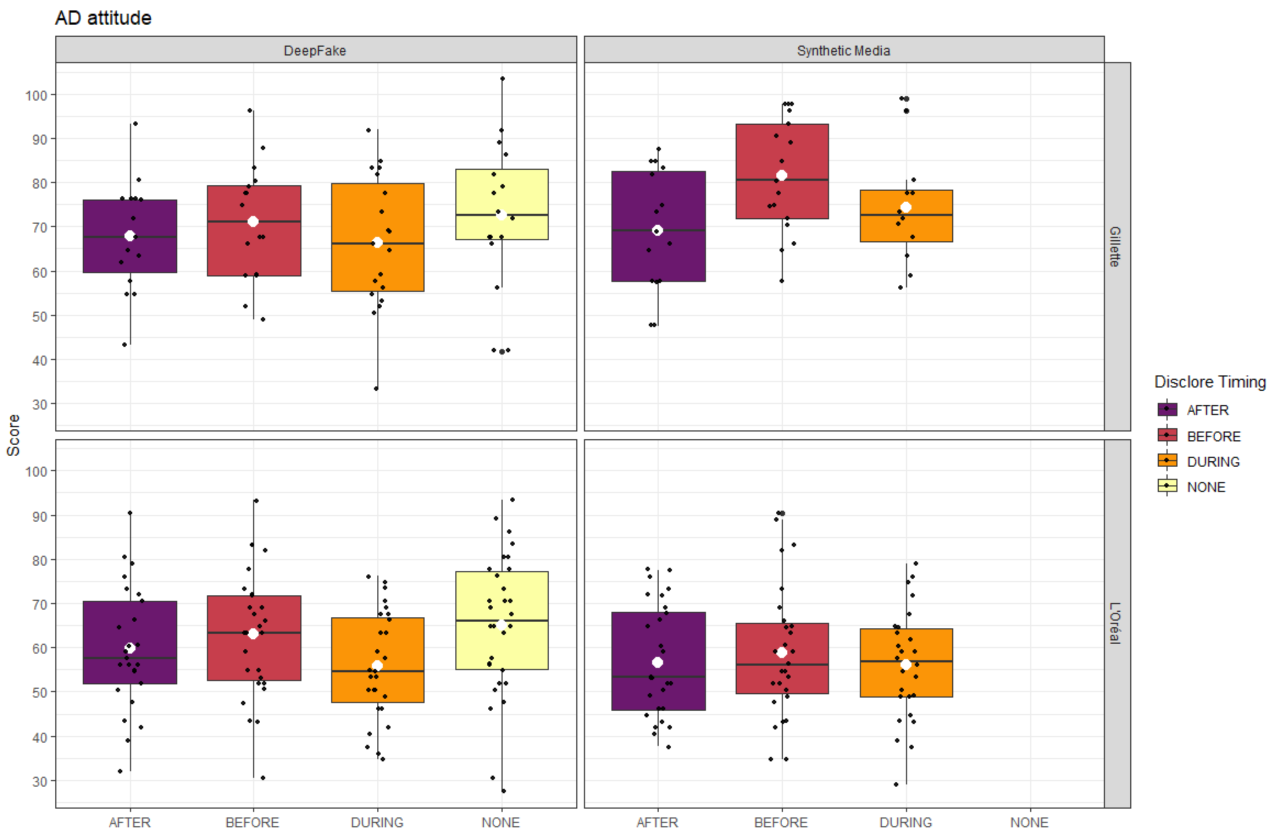


Figure 72 | Exposition 1 overview

From this plot, it's possible to notice how the average mean for the "Before" scenario is always higher than the rest of them as well as the "None" scenario. From this graph, we can also notice the presence of some outliers in scenarios N_DF_GIL and D_SM_GIL.

Exposition 2

The first thing is to check how many participants were familiar with the term "Deepfake" and "Synthetic Media". The possibility is that people were not familiar enough with the term synthetic media and could not react to it when displayed in the disclosure.
To check this, we plot the summary of the results:

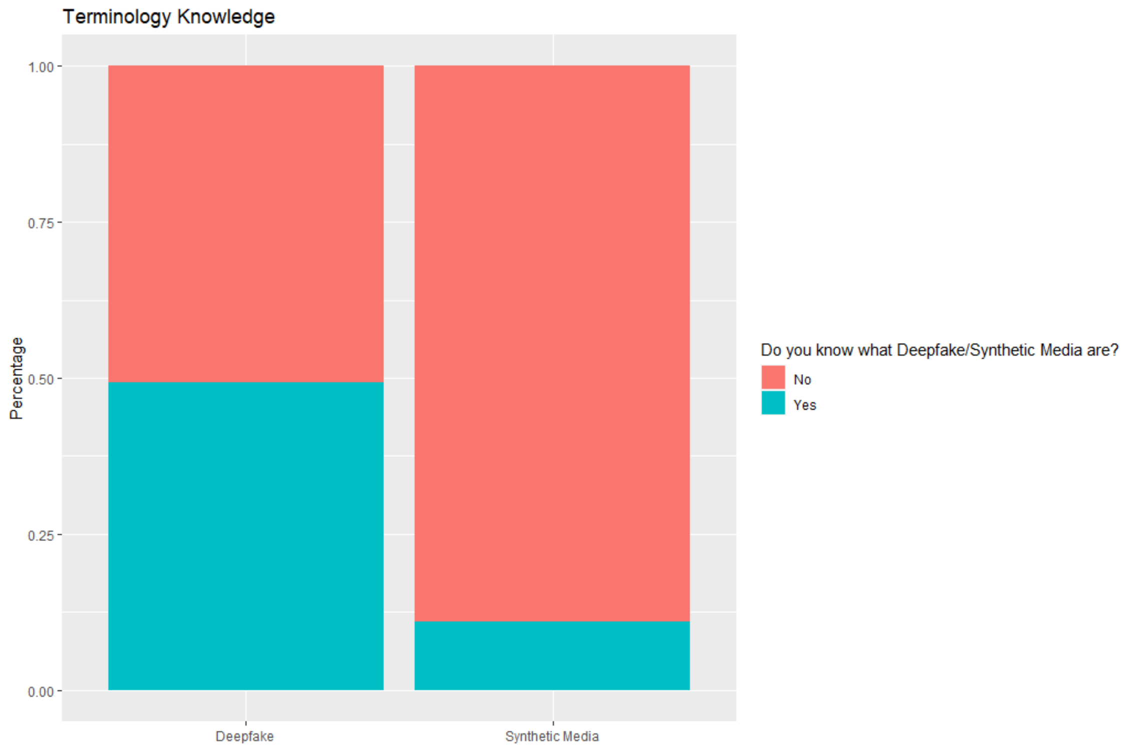


Figure 73 | Knowledge plot

As expected, respondents were more familiar with the term "Deepfake" than "Synthetic Media". In detail, 50.6% knew what deepfakes were, opposed to the 49.4% who didn't. We can't say the same for Synthetic Media where only 11% of the sample knew what they were opposed to a gigantic 89% who had no idea what they were. It could be plausible then, that the TYPE_1 interaction with DISCLOSURE_1 didn't take place because of a lack of knowledge.
Exposition 2 gets rid of this problem by making participants pass through a page of information about the technology as shown previously in an attempt to fix the knowledge asymmetry.
We now proceed similarly to Exposition 1 for finding the best-fit model.

Testing different types of ANOVA (exposition 2)

The models analyzed are:

One-way ANOVA: AD_TRUST_2 = DISCLOSURE_2

Two-way ANOVA: AD_TRUST_2 = DISCLOSURE_2 * TYPE_2

Two-way ANOVA + blocking variable: AD_TRUST_2 = DISCLOSURE_2 * TYPE_2 + BRAND_2

Three-way ANOVA: AD_TRUST_2 = DISCLOSURE_2 * TYPE_2 * value

Three-way ANOVA + blocking variable: AD_TRUST_2 = DISCLOSURE_2 * TYPE_2 * value + BRAND_2

Four-way ANOVA: AD_TRUST_2 = DISCLOSURE_2 * TYPE_2 * value * BRAND_2

"value" = previous knowledge of deepfake/synthetic media

Best-fit model comparison through Akaike Information Criterion (AIC) test.

```
Model selection based on AICc:

                 K    AICc Delta_AICc AICcWt Cum.Wt       LL
2W_BLOCK_ANOVA   8 2443.20       0.00   0.58   0.58 -1213.35
1W_ANOVA         4 2444.03       0.84   0.38   0.97 -1217.95
2W_ANOVA         7 2450.12       6.93   0.02   0.99 -1217.87
3W_BLOCK_ANOVA  14 2450.57       7.38   0.01   1.00 -1210.55
3W_ANOVA        13 2459.81      16.62   0.00   1.00 -1216.27
4W_ANOVA        25 2461.98      18.78   0.00   1.00 -1203.62
```

Figure 74 | AIC test, exposition 2

Exactly like for exposition 1, the two-way ANOVA + blocking variable is the best-fit one of the 6 tested. AICc score = 2443.20, slightly better than before. AICcWt = 0.58, much less than before. The result of the chosen model is shown below:

```
                     Df Sum Sq Mean Sq F value  Pr(>F)   
DISCLOSURE_2          2   1400   700.0   3.583 0.02901 * 
TYPE_2                1     15    15.2   0.078 0.78042   
BRAND_2               1   1721  1721.1   8.811 0.00324 **
DISCLOSURE_2:TYPE_2   2     45    22.5   0.115 0.89126   
Residuals           293  57234   195.3                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 75 | 2W_BLOCK_ANOVA, exposition 2

BRAND_2 still is the strongest variable with a p-value < .01. The second most important variable is again DISCLOSURE_2, with a p-value < .05.

Testing the model for homoscedasticity (exposition 2)

Checking if the model fits homoscedasticity by looking at the diagnostic plots.

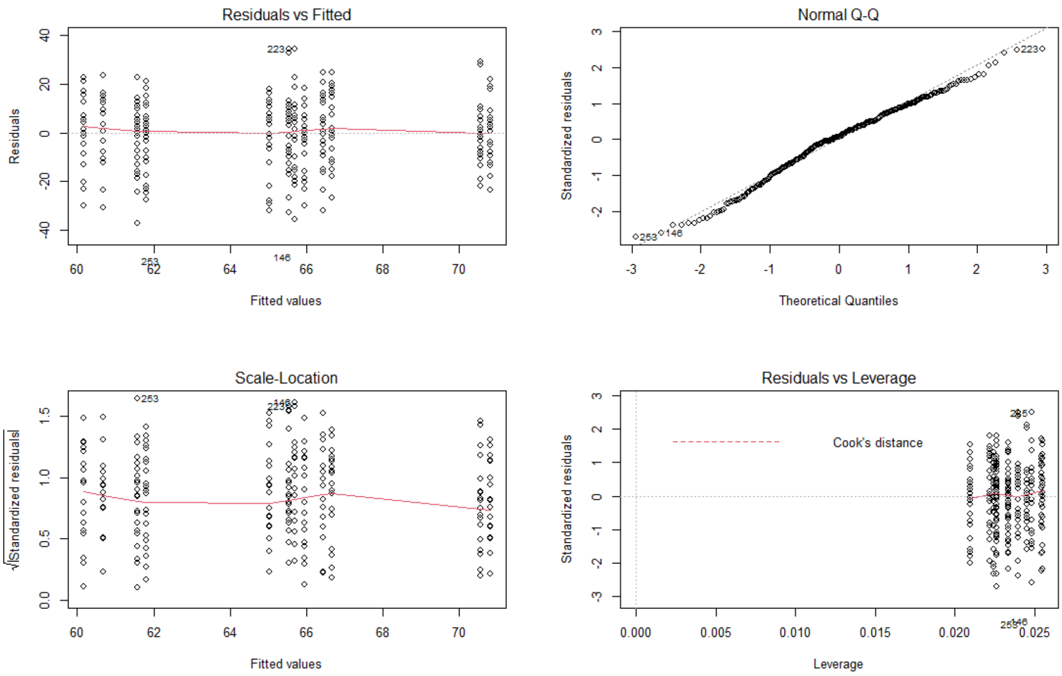


Figure 76 | Diagnostic plot model exposition 1

We don’t notice any big bias in the diagnostic plot. The model fits the assumption of homoscedasticity.

Post-hoc test (exposition 2)

Likewise, we proceed to run Turkey’s Honestly Significant Difference (Turkey’s HSD) post-hoc test.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = ADTRUST_2 ~ DISCLOSURE_2 * TYPE_2 + BRAND_2, data = MAIN_TEST)

$DISCLOSURE_2
           diff         lwr       upr      p adj
B-A  4.66424040  0.01687093 9.3116099 0.0489436
D-A  0.09587635 -4.52714034 4.7188930 0.9986850
D-B -4.56836405 -9.27202793 0.1352998 0.0590680

$TYPE_2
           diff       lwr      upr      p adj
SM-DF -0.450116 -3.627456 2.727224 0.7805883

$BRAND_2
            diff      lwr      upr      p adj
TCC-JAG 4.781925 1.603171 7.960679 0.0033206

$`DISCLOSURE_2:TYPE_2`
                diff        lwr       upr      p adj
B:DF-A:DF  3.7521844  -4.212087 11.716456 0.7557919
D:DF-A:DF -0.2519307  -8.006990  7.503128 0.9999990
A:SM-A:DF -1.3049931  -9.181135  6.571149 0.9969686
B:SM-A:DF  4.3287591  -3.505619 12.163138 0.6090977
D:SM-A:DF -0.8144308  -8.778702  7.149841 0.9997065
D:DF-B:DF -4.0041151 -12.073552  4.065322 0.7127431
A:SM-B:DF -5.0571775 -13.243048  3.128693 0.4853425
B:SM-B:DF  0.5765747  -7.569121  8.722270 0.9999522
D:SM-B:DF -4.5666152 -12.837316  3.704086 0.6098361
A:SM-D:DF -1.0530624  -9.035531  6.929406 0.9989843
B:SM-D:DF  4.5806898  -3.360575 12.521954 0.5629229
D:SM-D:DF -0.5625000  -8.631937  7.506937 0.9999557
B:SM-A:SM  5.6337522  -2.425798 13.693302 0.3417970
D:SM-A:SM  0.4905624  -7.695309  8.676433 0.9999791
D:SM-B:SM -5.1431899 -13.288886  3.002506 0.4600604
```

Figure 77 | Turkey’s HSD test for exposition 2

The only significant difference emerging from this test is between “After” and “Before” for DISCLOSURE_2, p < .05. Just outside the significance is the difference between “During and “Before” where .05 < p-value < .06.
No significant differences between TYPE_2 nor the interaction between TYPE_2 and DISCLOSURE_2, suggesting there is no moderating effect again. We then plot the result:

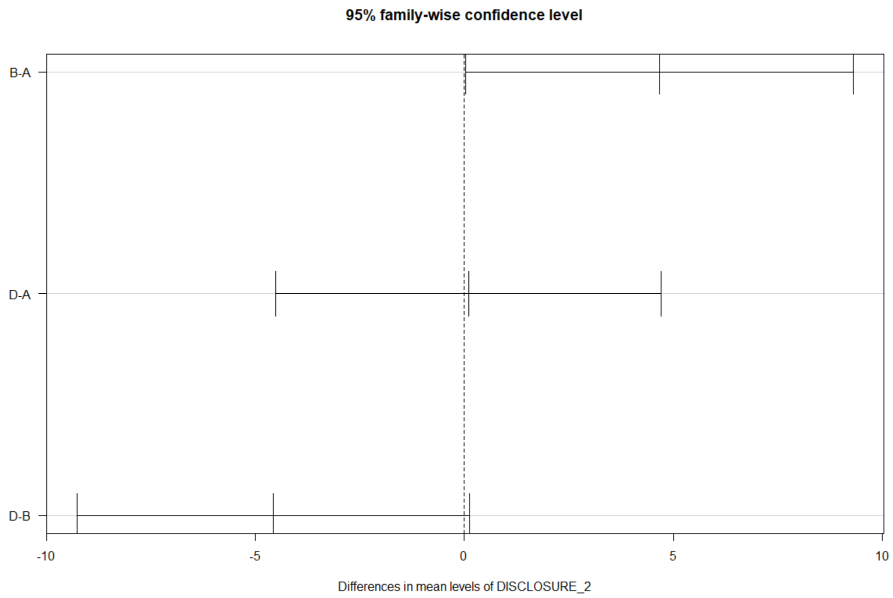
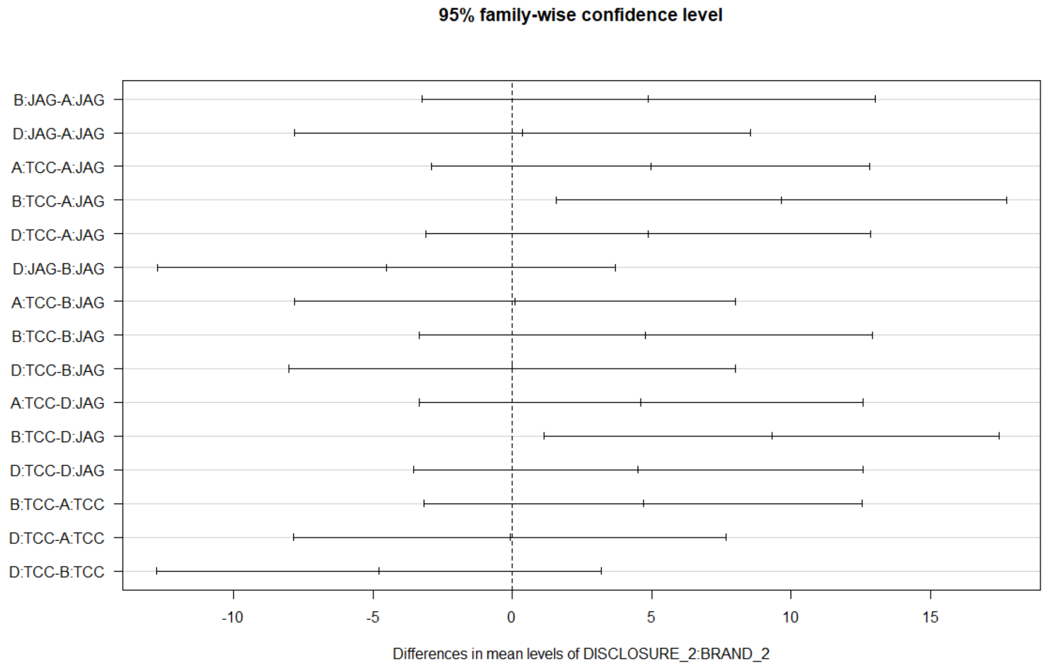


Figure 78 | the plotted result

What could be useful this time, is a plot of every single scenario divided by the brand to try and see if we can try and find some more insights.



From this plot, we notice that the difference between “Before” and “After” is significant for the Jaguar ad, while for TrueCaller there is a significant difference between “Before” and “During”. From this graph, and considering exposition 1 results, we can hypothesize disclosure only plays a difference if placed before, while it doesn’t make a big difference if placed during, or after the ad. We plot the interaction plot to have a clearer overview:

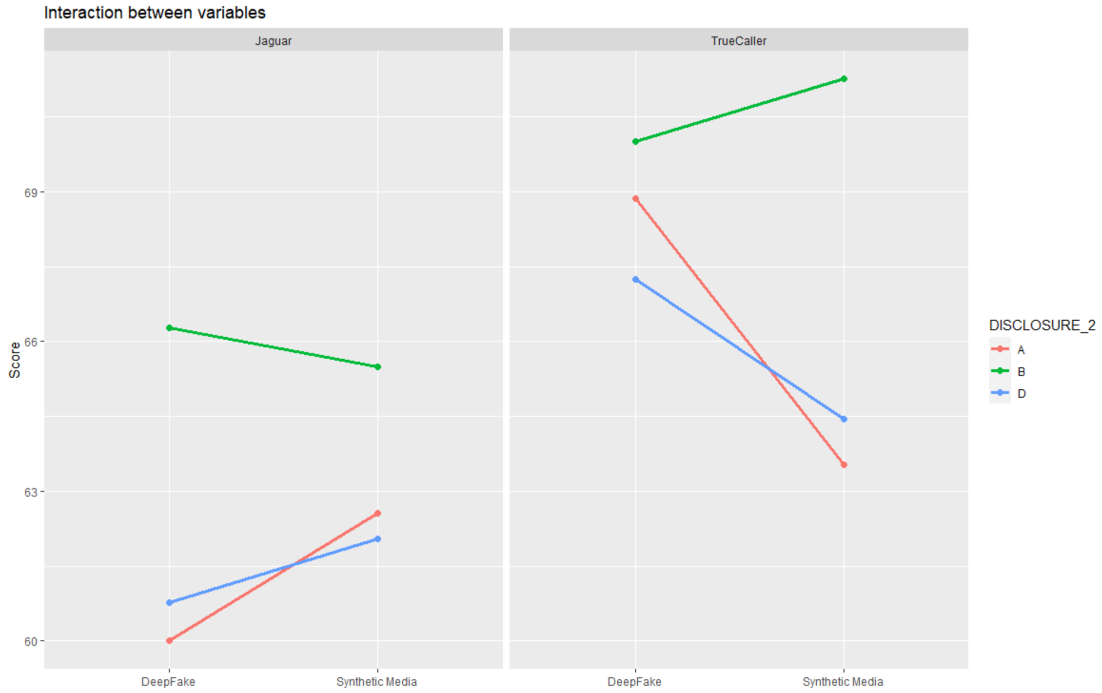


Figure 79 | Exposition_2 interaction plot

As we hypothesized, the “Before” scenario seems to have better overall results while “During” and “After” stay closer to each other’s values, in this case crossing each other. The variable TYPE_2 still doesn’t affect logically the outcome.

Overview plot for exposition 2

After validating the model, an overview plot for exposition 2 is created to summarize the differences.

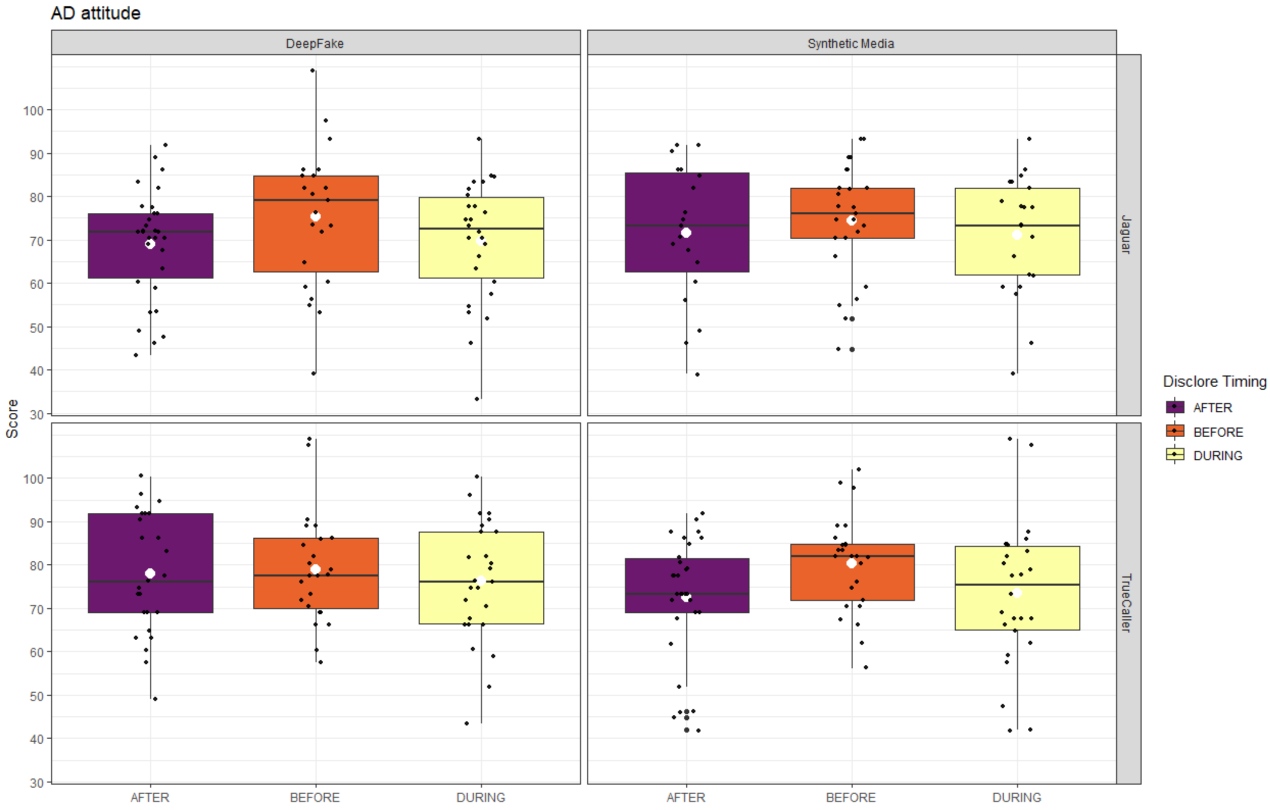


Figure 80 | Exposition 2 overview

As a final confirmation, “Before” performs better than all the other types of Disclosure placements. We can notice some outliers in B_SM_JAG and A_SM_TCC, but as we’ve seen before, they should not be responsible for too much bias for the model.

Managerial Implications

Deepfakes are an emerging technology that can drastically reduce marketing costs in many areas as described in the first chapter. From modeling to influencers, to advertising. AI-generated media can offer a new universe of solutions to marketers and managers in terms of creativity, accessibility, and brand building. Additionally, deepfakes and AI-generated media are creating new business opportunities. More and more start-ups are starting to emerge offering AI-generated content solutions to other brands and collaborating with marketing offices.
If not understood deeply in time, though, the risk is to invest money and resources into something that might backfire or that it's simply not clear how it works on consumers. On the other hand, the risk to stay behind and miss the train of the "competitive advantage" that its use could provide. Deepfakes still lie in a grey area that might soon be more defined. Disclosure regulations might come in the near future. Managers and marketers need to understand and study the best solution to avoid negative outcomes, maintain the proactive approach, and avoid being caught unprepared.

Future Research

This study showed different flaws for a lack of material. The analysis showed, as predictable, the type of ad showed was the most influential variable of the whole study. This study has been done using material found on the internet or the ones kindly shared by Chris Ume. The idea of creating an ad-hoc deepfake ad was out of the discussion for obvious hardware limitations', even though some attempts on Google Colab's[58] DeepFaceLab[59] project have been done. Unfortunately, other reasons linked to the time needed and Google Colab's time usage limits made this impossible. It would be interesting for future researchers to properly create a series of diversified ad-hoc ads to further investigate this topic precisely. Another problem that emerged during the data analysis phase was the small sample size. Unfortunately, the scenarios were too numerous, and too many respondents didn't pass the threshold for being considered in the study. Still, many outliers were observed. It would be interesting to improve this variable. Finally, a neuromarketing analysis with the aid of an ECG and Eye-tracking could further help in understanding the reasons and brain processes behind the exposition of a deepfake to a human.

Conclusions

This thesis covers an area of the unexplored field of deepfakes in advertising by studying the shifts in ads' attitudes derived by a disclosure message warning viewers of the use of deepfakes. Taking the overview analysis from chapter one and the harm such technology could cause; the use of it would likely be soon regulated. The technology is still in its infancy, but many countries are already moving towards regulation and many platforms are banning them if not respecting certain parameters or what we could define as "fair use". This study wants to provide a preliminary insight for managers and marketers on how to approach such a scenario by providing an analysis of the response consumers could have for different types of disclosure placements and terminology used. The results demonstrated that a disclosure message does affect the overall attitude towards the ad, but it also showed, contrarily to what was hypothesized, that placing it before the ad impacts in a much smaller magnitude compared to placing it during, or after it. Lastly, the difference in the terminology used doesn't seem to play a significant role.

Appendix

```
library(readr)
library(tidyverse)
library(devtools)
library(tidyverse)
library(hrbrthemes)
library(viridis)
library(xlsx)
library(psych)
library(car)
library(AICcmodavg)
library(rgeolocate)
library(scales)
library(dplyr)
library(ggplot2)
require(ggbiplot)
library(rworldmap)

#creating dataframe from .csv

Full_Dataset<-
read_delim("~/Tesimagistrale/Research/Dataset/Updated_Data_Gathering_Final_February+7,+20
22_11.56.csv",
                    ";", escape_double = FALSE, trim_ws = TRUE)


# plot data on world map

worldmap <- getMap(resolution = "coarse")
plot(worldmap, xlim = c(-150, 150), ylim = c(-80, 100),
   asp = 1, border = "darkgray", bg = "aliceblue", col = "black", fill = T)

# add points

points(Full_Dataset$LONGITUDE, Full_Dataset$LATITUDE,
     col = "red", cex = 1, pch = 20)
title(main = "Respondents Map Overview")

#Substituting likert scale to numerical values for the first exposition

Full_Dataset$AD_BORING_1 <- as.integer(plyr::mapvalues(Full_Dataset$AD_BORING_1,
                                      from = c("Strongly disagree",
                                            "Disagree",
                                            "Somewhat disagree",
                                            "Neither agree nor disagree",
                                            "Somewhat agree",
                                            "Agree",
                                            "Strongly agree"),
                                      to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_IRRITATING_1 <- as.integer(plyr::mapvalues(Full_Dataset$AD_IRRITATING_1,
                                         from = c("Strongly disagree",
                                               "Disagree",
                                               "Somewhat disagree",
```

```
                                                     "Neither agree nor disagree",
                                                     "Somewhat agree",
                                                     "Agree",
                                                     "Strongly agree"),
                                              to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_DISTURBING_1                                                                                            <-
as.integer(plyr::mapvalues(Full_Dataset$AD_DISTURBING_1,
                                              from = c("Strongly disagree",
                                                     "Disagree",
                                                     "Somewhat disagree",
                                                     "Neither agree nor disagree",
                                                     "Somewhat agree",
                                                     "Agree",
                                                     "Strongly agree"),
                                              to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_CREDIBLE_1 <- as.integer(plyr::mapvalues(Full_Dataset$AD_CREDIBLE_1,
                                             from = c("Strongly disagree",
                                                    "Disagree",
                                                    "Somewhat disagree",
                                                    "Neither agree nor disagree",
                                                    "Somewhat agree",
                                                    "Agree",
                                                    "Strongly agree"),
                                             to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_GOOD_1 <- as.integer(plyr::mapvalues(Full_Dataset$AD_GOOD_1,
                                          from = c("Strongly disagree",
                                                 "Disagree",
                                                 "Somewhat disagree",
                                                 "Neither agree nor disagree",
                                                 "Somewhat agree",
                                                 "Agree",
                                                 "Strongly agree"),
                                          to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_HONEST_1 <- as.integer(plyr::mapvalues(Full_Dataset$AD_HONEST_1,
                                           from = c("Strongly disagree",
                                                  "Disagree",
                                                  "Somewhat disagree",
                                                  "Neither agree nor disagree",
                                                  "Somewhat agree",
                                                  "Agree",
                                                  "Strongly agree"),
                                           to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_TRUTHFUL_1 <- as.integer(plyr::mapvalues(Full_Dataset$AD_TRUTHFUL_1,
                                             from = c("Strongly disagree",
                                                    "Disagree",
                                                    "Somewhat disagree",
                                                    "Neither agree nor disagree",
                                                    "Somewhat agree",
                                                    "Agree",
                                                    "Strongly agree"),
                                             to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_LIKEABLE_1 <- as.integer(plyr::mapvalues(Full_Dataset$AD_LIKEABLE_1,
                                             from = c("Strongly disagree",
                                                    "Disagree",
                                                    "Somewhat disagree",
```

```
                                                "Neither agree nor disagree",
                                                "Somewhat agree",
                                                "Agree",
                                                "Strongly agree"),
                                           to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_ENJOYABLE_1 <- as.integer(plyr::mapvalues(Full_Dataset$AD_ENJOYABLE_1, from = c("Strongly disagree", "Disagree", "Somewhat disagree", "Neither agree nor disagree", "Somewhat agree", "Agree", "Strongly agree"), to = c(1,2,3,4,5,6,7)))
Full_Dataset$LIKE_1 <- as.integer(plyr::mapvalues(Full_Dataset$LIKE_1, from = c("Dislike a great deal", "Dislike a moderate amount", "Dislike a little", "Neither like nor dislike", "Like a little", "Like a moderate amount", "Like a great deal"), to = c(1,2,3,4,5,6,7)))

#Doing the same for the second exposition

Full_Dataset$AD_BORING_2 <- as.integer(plyr::mapvalues(Full_Dataset$AD_BORING_2, from = c("Strongly disagree", "Disagree", "Somewhat disagree", "Neither agree nor disagree", "Somewhat agree", "Agree", "Strongly agree"), to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_IRRITATING_2 <- as.integer(plyr::mapvalues(Full_Dataset$AD_IRRITATING_2, from = c("Strongly disagree", "Disagree", "Somewhat disagree", "Neither agree nor disagree", "Somewhat agree", "Agree", "Strongly agree"), to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_DISTURBING_2 <- as.integer(plyr::mapvalues(Full_Dataset$AD_DISTURBING_2, from = c("Strongly disagree", "Disagree", "Somewhat disagree", "Neither agree nor disagree", "Somewhat agree", "Agree", "Strongly agree"), to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_CREDIBLE_2 <- as.integer(plyr::mapvalues(Full_Dataset$AD_CREDIBLE_2, from = c("Strongly disagree", "Disagree", "Somewhat disagree", "Neither agree nor disagree", "Somewhat agree", "Agree", "Strongly agree"), to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_GOOD_2 <- as.integer(plyr::mapvalues(Full_Dataset$AD_GOOD_2, from = c("Strongly disagree", "Disagree", "Somewhat disagree", "Neither agree nor disagree", "Somewhat agree", "Agree", "Strongly agree"), to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_HONEST_2 <- as.integer(plyr::mapvalues(Full_Dataset$AD_HONEST_2, from = c("Strongly disagree", "Disagree", "Somewhat disagree", "Neither agree nor disagree", "Somewhat agree", "Agree", "Strongly agree"), to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_TRUTHFUL_2 <- as.integer(plyr::mapvalues(Full_Dataset$AD_TRUTHFUL_2, from = c("Strongly disagree", "Disagree", "Somewhat disagree", "Neither agree nor disagree", "Somewhat agree", "Agree", "Strongly agree"), to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_LIKEABLE_2 <- as.integer(plyr::mapvalues(Full_Dataset$AD_LIKEABLE_2, from = c("Strongly disagree", "Disagree", "Somewhat disagree", "Neither agree nor disagree", "Somewhat agree", "Agree", "Strongly agree"), to = c(1,2,3,4,5,6,7)))
Full_Dataset$AD_ENJOYABLE_2 <- as.integer(plyr::mapvalues(Full_Dataset$AD_ENJOYABLE_2, from = c("Strongly disagree", "Disagree", "Somewhat disagree", "Neither agree nor disagree", "Somewhat agree", "Agree", "Strongly agree"), to = c(1,2,3,4,5,6,7)))
Full_Dataset$LIKE_2 <- as.integer(plyr::mapvalues(Full_Dataset$LIKE_2, from = c("Dislike a great deal", "Dislike a moderate amount", "Dislike a little", "Neither like nor dislike", "Like a little", "Like a moderate amount", "Like a great deal"), to = c(1,2,3,4,5,6,7)))

#Deleting extra columns

Full_Dataset[67:81] <- list(NULL)

#Creating Education Levels abbreviations

Full_Dataset <- Full_Dataset %>%
```

```
  mutate(EDUCATION = if_else(EDUCATION == "2nd Level Master", "2ndL Master", EDUCATION)) %>%
  mutate(EDUCATION = if_else(EDUCATION == "High school graduate", "HighSchool", EDUCATION)) %>%
  mutate(EDUCATION = if_else(EDUCATION == "Bachelor Degree", "Bachelor", EDUCATION)) %>%
  mutate(EDUCATION = if_else(EDUCATION == "Less than high school", "No HighSchool", EDUCATION)) %>%
  mutate(EDUCATION = if_else(EDUCATION == "Postgraduate/Master/5 year Degree", "Postgraduate", EDUCATION))

#creating subset for every different scenario

###Creating PRE_TEST subset from FULL_DATASET

PRE_TEST <- Full_Dataset
PRE_TEST[1:3] <- list(NULL)
PRE_TEST[2] <- NULL
PRE_TEST <- pivot_longer(PRE_TEST, cols = 2:5, names_to = c("nothign", "noge", "N_DF", "BRAND"), names_sep = "_", values_drop_na = TRUE)
PRE_TEST[3:52] <- list(NULL)
PRE_TEST[6:10] <- NULL
PRE_TEST[8] <- NULL

#Creating Main_Test subset from FULL_DATASET

MAIN_TEST <- Full_Dataset
MAIN_TEST[1:3] <- NULL
MAIN_TEST[2:7] <- NULL
MAIN_TEST <- pivot_longer(MAIN_TEST,
                    cols = B_DF_LOR:N_DF_GILL,
                    names_to = c("DISCLOSURE_1", "TYPE_1", "BRAND_1"),
                    names_sep = "_",
                    values_drop_na = TRUE)
MAIN_TEST[13] <- NULL
MAIN_TEST[14] <- NULL
MAIN_TEST[45] <- NULL
MAIN_TEST <- pivot_longer(MAIN_TEST,
                    cols = B_DF_JAG:A_SM_TCC,
                    names_to = c("DISCLOSURE_2", "TYPE_2", "BRAND_2"),
                    names_sep = "_",
                    values_drop_na = TRUE)
MAIN_TEST[36] <- NULL
MAIN_TEST[27:29] <- NULL

#fixing the mispelled "GILL" in BRAND + convert everything in factor

MAIN_TEST <- MAIN_TEST %>%
  mutate(BRAND_1 = if_else(BRAND_1 == "GILL", "GIL", BRAND_1)) %>%
  mutate_if(is.character, factor)

PRE_TEST <- PRE_TEST %>%
  mutate(BRAND = if_else(BRAND == "GILL", "GIL", BRAND)) %>%
  mutate_if(is.character, factor)
```

```
#creating one column for KNOWLEDGE

MAIN_TEST <- pivot_longer(MAIN_TEST, cols = DF_KNOWLEDGE:SM_KNOWLEDGE, names_to
= "Knowledge", values_drop_na = TRUE)

#PRE_TEST ANALYSIS

#demographics main test

PT_Demographics <- PRE_TEST[3:5]
PT_Demographics$Gender <- PRE_TEST$Gender
summary(PT_Demographics)

#gender

Gender_pre <- PRE_TEST %>%
  group_by(Gender) %>% # Variable to be transformed
  dplyr::count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))

ggplot(Gender_pre, aes(x = "", y = perc, fill = Gender)) +
  geom_col() +
  geom_text(aes(x=1.6, label = labels),
            position = position_stack(vjust = 0.5))+
  coord_polar(theta = "y")+
  guides(fill = guide_legend(title = "Gender"))+
  scale_fill_viridis_d(option = "A", begin = 0, end = 0.8, drop = F)+
  theme_bw()+
  theme_void()

#Education

Education_pre <- PRE_TEST %>%
  group_by(EDUCATION) %>% # Variable to be transformed
  dplyr::count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))

ggplot(Education_pre, aes(x = "", y = perc, fill = EDUCATION)) +
  geom_col() +
  geom_text(aes(x=1.55, label = labels),
            position = position_stack(vjust = 0.5))+
  coord_polar(theta = "y")+
  guides(fill = guide_legend(title = "Education"))+
  scale_fill_viridis_d(option = "A", begin = 0, end = 0.85, drop = F)+
  theme_bw()+
  theme_void()

#Age
```

```
Age_pre <- PRE_TEST %>%
  group_by(AGE) %>% # Variable to be transformed
  dplyr::count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))

ggplot(Age_pre, aes(x = "", y = perc, fill = AGE)) +
  geom_col() +
  geom_text(aes(x=1.55, label = labels),
            position = position_stack(vjust = 0.5))+
  coord_polar(theta = "y")+
  guides(fill = guide_legend(title = "Age"))+
  scale_fill_viridis_d(option = "A", begin = 0, end = 0.85, drop = F)+
  theme_bw()+
  theme_void()

#Employment

Employment_pre <- PRE_TEST %>%
  group_by(EMPLOYMENT) %>% # Variable to be transformed
  dplyr::count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))

ggplot(Employment_pre, aes(x = "", y = perc, fill = EMPLOYMENT)) +
  geom_col() +
  geom_text(aes(x=1.55, label = labels),
            position = position_stack(vjust = 0.5))+
  coord_polar(theta = "y")+
  guides(fill = guide_legend(title = "Employment"))+
  scale_fill_viridis_d(option = "A", begin = 0, end = 0.85, drop = F)+
  theme_bw()+
  theme_void()

#mutate variables into factor

PRE_TEST <- PRE_TEST %>%
  mutate(across(everything(), factor))

#preparing labels for plotting

likert_breaks <- c("Strongly Disagree", "Somewhat Disagree", "Slightly Disagree", "Neutral", "Slightly
Agree", "Somewhat Agree", "Strongly Agree")

score_breaks <- c("10","20","30","40","50","60","70","80","90","100")

labeller_spot <- c(
  GIL = "Gillette",
  LOR = "L'Oréal",
  JAG = "Jaguar",
  TCC = "TrueCaller"
```

```
)

labeller_pre_test_cond <- c(
  DF = "DeepFake",
  N = "Not DeepFake"
)

#PLOTTING RESULTS

PRE_TEST %>%
  ggplot(aes(x = N_DF, y = 100, fill = REAL_or_FAKE)) +
  geom_bar(position = "fill", stat = "identity")+
  facet_grid(~BRAND, labeller = labeller(
    BRAND = labeller_spot
  ),
  switch = "y")+
  scale_fill_viridis_d(option = "A", begin = 0, end = 0.8, drop = F)+
  ggtitle("Deepfake recognition by Type of Ad")+
  xlab("")+
  ylab("Percentage")+
  labs(fill = "REAL OR FAKE?")+
  theme_bw()+
  theme(axis.text = element_text(size = 10))+
  scale_x_discrete(labels = labeller_pre_test_cond)

SUMMARY_PRE_TEST <- data.frame(PRE_TEST$REAL_or_FAKE)
SUMMARY_PRE_TEST$N_DF <- PRE_TEST$N_DF
SUMMARY_PRE_TEST$BRAND <- PRE_TEST$BRAND
table(SUMMARY_PRE_TEST)

#________________________PRE_TEST ANALYSED________________________________


#_______________BEGINNING OF MAIN_TEST_____________________________

#demographics main test

MT_Demographics <- MAIN_TEST[22:24]
MT_Demographics$Gender <- MAIN_TEST$Gender
summary(MT_Demographics)

#gender

Gender_main <- MAIN_TEST %>%
  group_by(Gender) %>% # Variable to be transformed
  dplyr::count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))

ggplot(Gender_main, aes(x = "", y = perc, fill = Gender)) +
  geom_col() +
  geom_text(aes(x=1.6, label = labels),
            position = position_stack(vjust = 0.5))+
```

```
  coord_polar(theta = "y")+
  guides(fill = guide_legend(title = "Gender"))+
  scale_fill_viridis_d(option = "A", begin = 0, end = 0.8, drop = F)+
  theme_bw()+
  theme_void()

#Education


Education_main <- MAIN_TEST %>%
  group_by(EDUCATION) %>% # Variable to be transformed
  dplyr::count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))

ggplot(Education_main, aes(x = "", y = perc, fill = EDUCATION)) +
  geom_col() +
  geom_text(aes(x=1.55, label = labels),
            position = position_stack(vjust = 0.5))+
  coord_polar(theta = "y")+
  guides(fill = guide_legend(title = "Education"))+
  scale_fill_viridis_d(option = "A", begin = 0, end = 0.85, drop = F)+
  theme_bw()+
  theme_void()

#Age

Age_main <- MAIN_TEST %>%
  group_by(AGE) %>% # Variable to be transformed
  dplyr::count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
  mutate(labels = scales::percent(perc))

ggplot(Age_main, aes(x = "", y = perc, fill = AGE)) +
  geom_col() +
  geom_text(aes(x=1.55, label = labels),
            position = position_stack(vjust = 0.5))+
  coord_polar(theta = "y")+
  guides(fill = guide_legend(title = "Age"))+
  scale_fill_viridis_d(option = "A", begin = 0, end = 0.85, drop = F)+
  theme_bw()+
  theme_void()

#Employment

Employment_main <- MAIN_TEST %>%
  group_by(EMPLOYMENT) %>% # Variable to be transformed
  dplyr::count() %>%
  ungroup() %>%
  mutate(perc = `n` / sum(`n`)) %>%
  arrange(perc) %>%
```

```
  mutate(labels = scales::percent(perc))

ggplot(Employment_main, aes(x = "", y = perc, fill = EMPLOYMENT)) +
  geom_col() +
  geom_text(aes(x=1.55, label = labels),
            position = position_stack(vjust = 0.5))+
  coord_polar(theta = "y")+
  guides(fill = guide_legend(title = "Employment"))+
  scale_fill_viridis_d(option = "A", begin = 0, end = 0.85, drop = F)+
  theme_bw()+
  theme_void()


#creating data.frame of likert to be analyzed for pca and cronbach's alpha

reliability_1 <- MAIN_TEST[2:11]
reliability_2 <- MAIN_TEST[12:21]

#pca test for 1 and 2

pca_1 <- princomp(reliability_1, cor = T)
part_pca_1 <- pca_1$sdev^2/sum(pca_1$sdev^2)*100
print(part_pca_1)
print(cumsum(part_pca_1))

ggbiplot(pca_1)

pca_2 <- princomp(reliability_2, cor = T)
part_pca_2 <- pca_2$sdev^2/sum(pca_2$sdev^2)*100
print(part_pca_2)
print(cumsum(part_pca_2))

ggbiplot(pca_2)


#checking scales with cronbach's alpha, I use check.keys = true as the first three items are negatively
correlated with the total scale.

#reverse.coding negative related items

#recode inverted

#Checking Cronbach's alpha

reliability_1$AD_BORING_1 <- as.list(reliability_1$AD_BORING_1)
reliability_1$AD_BORING_1 <- as.integer(recode(reliability_1$AD_BORING_1, '1=7; 2=6; 3=5; 4=4;
5=3; 6=2; 7=1'))

reliability_1$AD_IRRITATING_1 <- as.list(reliability_1$AD_IRRITATING_1)
reliability_1$AD_IRRITATING_1 <- as.integer(recode(reliability_1$AD_IRRITATING_1, '1=7; 2=6;
3=5; 4=4; 5=3; 6=2; 7=1'))

reliability_1$AD_DISTURBING_1 <- as.list(reliability_1$AD_DISTURBING_1)
reliability_1$AD_DISTURBING_1 <- as.integer(recode(reliability_1$AD_DISTURBING_1, '1=7;
2=6; 3=5; 4=4; 5=3; 6=2; 7=1'))
```

```
psych::alpha(reliability_1)

reliability_2$AD_BORING_2 <- as.list(reliability_2$AD_BORING_2)
reliability_2$AD_BORING_2 <- as.integer(recode(reliability_2$AD_BORING_2, '1=7; 2=6; 3=5; 4=4;
5=3; 6=2; 7=1'))


reliability_2$AD_IRRITATING_2 <- as.list(reliability_2$AD_IRRITATING_2)
reliability_2$AD_IRRITATING_2 <- as.integer(recode(reliability_2$AD_IRRITATING_2, '1=7; 2=6;
3=5; 4=4; 5=3; 6=2; 7=1'))

reliability_2$AD_DISTURBING_2 <- as.list(reliability_2$AD_DISTURBING_2)
reliability_2$AD_DISTURBING_2 <- as.integer(recode(reliability_2$AD_DISTURBING_2, '1=7;
2=6; 3=5; 4=4; 5=3; 6=2; 7=1'))


psych::alpha(reliability_2, check.keys = TRUE)

#test passed = Cronbach alpha = 0.9 > 0.7 for exposition 1 and 0.88 > 0.7 for exposition 2.

#Creating labels

Disclosure_types <- c("AFTER", "BEFORE", "DURING", "NONE")

Labeller_terminology <- c(
  DF = "DeepFake",
  SM = "Synthetic Media"
)

#creating means for overview of attitude

MAIN_TEST$ADTRUST_1 <- apply(reliability_1, 1,mean)
MAIN_TEST$ADTRUST_2 <- apply(reliability_2, 1,mean)

#transforming ADTRUST likert scale mean into a 100 score

MAIN_TEST$ADTRUST_1 <- MAIN_TEST$ADTRUST_1*100/7
MAIN_TEST$ADTRUST_2 <- MAIN_TEST$ADTRUST_2*100/7


#Analyzing the model

#ANOVA Analysis for exposition 1

one_way_ANOVA_1 <- aov(ADTRUST_1 ~ DISCLOSURE_1,
                      data = MAIN_TEST)
summary(one_way_ANOVA_1)

two_way_ANOVA_1 <- aov(ADTRUST_1 ~ DISCLOSURE_1*TYPE_1,
                      data = MAIN_TEST)
summary(two_way_ANOVA_1)

blocking_ANOVA_1 <- aov(ADTRUST_1 ~ DISCLOSURE_1*TYPE_1 + BRAND_1,
```

```
                    data = MAIN_TEST)
summary(blocking_ANOVA_1)

value_ANOVA_1 <- aov(ADTRUST_1 ~ DISCLOSURE_1*TYPE_1*value,
                  data = MAIN_TEST)
summary(value_ANOVA_1)

value_blocking_ANOVA_1 <- aov(ADTRUST_1 ~ DISCLOSURE_1*TYPE_1*value + BRAND_1,
                         data = MAIN_TEST)
summary(value_blocking_ANOVA_1)

value_brand_ANOVA_1 <- aov(ADTRUST_1 ~ DISCLOSURE_1*TYPE_1*value*BRAND_1,
                      data = MAIN_TEST)
summary(value_brand_ANOVA_1)

#finding the best-fit model for exposition 1

model.set_1 <- list(one_way_ANOVA_1, two_way_ANOVA_1, blocking_ANOVA_1,
value_ANOVA_1, value_blocking_ANOVA_1, value_brand_ANOVA_1)
model.names_1 <- c("1W_ANOVA", "2W_ANOVA", "2W_BLOCK_ANOVA", "3W_ANOVA",
"3W_BLOCK_ANOVA", "4W_ANOVA")

aictab(model.set_1, modnames = model.names_1)

#testing the model for homoscedasticity in exposition 1

par(mfrow=c(2,2))
plot(blocking_ANOVA_1)
par(mfrow=c(1,1))

#post-hoc test for exposition 1

tukey.blockinganova_1 <-TukeyHSD(blocking_ANOVA_1)

tukey.blockinganova_1

#Find the groupwise differences for exposition 1

tukey.plot.aov<-aov(ADTRUST_1 ~ DISCLOSURE_1*BRAND_1, data= MAIN_TEST)
tukey.plot.test<-TukeyHSD(tukey.plot.aov)
plot(tukey.plot.test, las = 1)
par(mar=c(5,6,4,1)+.15)

#plotting first exposition interaction between DISCLOSURE_1, TYPE_1 by BRAND_1

MAIN_TEST %>%
  ggplot(aes(x = TYPE_1, group = DISCLOSURE_1, color = DISCLOSURE_1, y = ADTRUST_1))+
  stat_summary(fun = mean, geom = "point", lwd = 2)+
  stat_summary(fun = mean, geom = "line", lwd = 1.1)+
  facet_grid(~BRAND_1, labeller = labeller(
    BRAND_1 = labeller_spot, switch = "y"))+
  xlab("")+
  ylab("Score")+
  labs(fill = "Disclosure Timing")+
  ggtitle("Interaction between variables")+
```

```
  scale_x_discrete(labels = Labeller_terminology)+
  scale_fill_discrete(name = "Disclosure Timing", labels = c("After", "Before", "During", "None"))

#plotting overview_1

MAIN_TEST %>%
  ggplot(aes(x = DISCLOSURE_1, y = ADTRUST_1, fill = DISCLOSURE_1)) +
  geom_boxplot(stat = "boxplot", position = "dodge") +
  stat_summary(fun = mean, geom = "point", shape = 20, size = 5, color = "white", fill = "red")+
  geom_jitter(color = "black", size = 1, alpha = 0.9, width = 0.1, height = 0.1)+
  theme_ipsum()+
  theme(
    legend.position = "none",
    plot.title = element_text(size = 11)
  )+
  ggtitle("AD attitude")+
  xlab("")+
  ylab("Score")+
  labs(fill = "Disclore Timing")+
  scale_x_discrete(labels = Disclosure_types)+
  scale_y_continuous("Score", labels = score_breaks, breaks = seq(1, 100, by = 10))+
  facet_grid(BRAND_1 ~ TYPE_1, labeller = labeller(TYPE_1 = Labeller_terminology, BRAND_1 =
labeller_spot)) +
  scale_fill_viridis_d(option = "B", begin = 0.3, end = 1, labels = Disclosure_types, drop = F) +
  theme_bw()

#Plot knowledge for DF and SM

MAIN_TEST %>%
  group_by(Knowledge, value) %>%
  dplyr::summarise(count = n())

Knowledge_labels <- c("Deepfake", "Synthetic Media")

MAIN_TEST %>%
  ggplot(aes(Knowledge, y= 100, fill = value))+
  ggtitle("Terminology Knowledge")+
  geom_bar(position = "fill", stat = "identity")+
  xlab("")+
  ylab("Percentage")+
  labs(fill = "Do you know what Deepfake/Synthetic Media are?")+
  scale_x_discrete(labels = Knowledge_labels)

#ANOVA Analysis for exposition 2 interaction

one_way_ANOVA_2 <- aov(ADTRUST_2 ~ DISCLOSURE_2,
                       data = MAIN_TEST)
summary(one_way_ANOVA_2)

two_way_ANOVA_2 <- aov(ADTRUST_2 ~ DISCLOSURE_2*TYPE_2,
                       data = MAIN_TEST)
summary(two_way_ANOVA_2)

blocking_ANOVA_2 <- aov(ADTRUST_2 ~ DISCLOSURE_2*TYPE_2 + BRAND_2,
                        data = MAIN_TEST)
```

```
summary(blocking_ANOVA_2)

value_ANOVA_2 <- aov(ADTRUST_2 ~ DISCLOSURE_2*TYPE_2*value,
                     data = MAIN_TEST)
summary(value_ANOVA_2)

value_blocking_ANOVA_2 <- aov(ADTRUST_2 ~ DISCLOSURE_2*TYPE_2*value + BRAND_2,
                              data = MAIN_TEST)
summary(value_blocking_ANOVA_2)

value_brand_ANOVA_2 <- aov(ADTRUST_2 ~ DISCLOSURE_2*TYPE_2*value*BRAND_2,
                           data = MAIN_TEST)
summary(value_brand_ANOVA_2)

#finding the best-fit model for exposition 2

model.set_2 <- list(one_way_ANOVA_2, two_way_ANOVA_2, blocking_ANOVA_2,
value_ANOVA_2, value_blocking_ANOVA_2, value_brand_ANOVA_2)
model.names_2 <- c("1W_ANOVA", "2W_ANOVA", "2W_BLOCK_ANOVA", "3W_ANOVA",
"3W_BLOCK_ANOVA", "4W_ANOVA")

aictab(model.set_2, modnames = model.names_2)

#testing the models for homoscedasticity for exposition 2

par(mfrow=c(2,2))
plot(blocking_ANOVA_2)
par(mfrow=c(1,1))

#post-hoc test for exposition 2

tukey.blockinganova_2 <-TukeyHSD(blocking_ANOVA_2)

tukey.blockinganova_2

#Find the groupwise differences for exposition 2

par(mar=c(5,9,6,1)+.2)
tukey.plot.aov_2 <-aov(ADTRUST_2 ~ DISCLOSURE_2*BRAND_2, data= MAIN_TEST)
tukey.plot.test_2 <-TukeyHSD(tukey.plot.aov_2)
plot(tukey.plot.test_2 , las = 1)

#plotting second exposition interaction between DISCLOSURE_2 and TYPE_2

MAIN_TEST %>%
  ggplot(aes(x = TYPE_2, group = DISCLOSURE_2, color = DISCLOSURE_2, y = ADTRUST_2))+
  stat_summary(fun = mean, geom = "point", lwd = 2)+
  stat_summary(fun = mean, geom = "line", lwd = 1.1)+
  facet_grid(~BRAND_2, labeller = labeller(
    BRAND_2 = labeller_spot, switch = "y"))+
  xlab("")+
  ylab("Score")+
  labs(fill = "Disclosure Timing")+
  ggtitle("Interaction between variables")+
  scale_x_discrete(labels = Labeller_terminology)+
```

```
  scale_fill_discrete(name = "Disclosure Timing", labels = c("After", "Before", "During", "None"))

#plotting overview 2

MAIN_TEST %>%
  ggplot(aes(x = DISCLOSURE_2, y = ADTRUST_2, fill = DISCLOSURE_2)) +
  geom_boxplot(stat = "boxplot", position = "dodge") +
  stat_summary(fun = mean, geom = "point", shape = 20, size = 5, color = "white", fill = "red")+
  geom_jitter(color = "black", size = 1, alpha = 0.9, width = 0.1, height = 0.1)+
  theme_ipsum()+
  theme(
    legend.position = "none",
    plot.title = element_text(size = 11)
  )+
  ggtitle("AD attitude")+
  xlab("")+
  ylab("Score")+
  labs(fill = "Disclore Timing")+
  scale_x_discrete(labels = Disclosure_types)+
  scale_y_continuous("Score", labels = score_breaks, breaks = seq(1, 100, by = 10))+
  facet_grid(BRAND_2 ~ TYPE_2, labeller = labeller(TYPE_2 = Labeller_terminology, BRAND_2 =
labeller_spot)) +
  scale_fill_viridis_d(option = "B", begin = 0.3, end = 1, labels = Disclosure_types, drop = F) +
  theme_bw()

write.xlsx(MAIN_TEST, file = "MAINTEST_2.xlsx")
```

References

AB-730. 2019. “Bill Text - AB-730 Elections: Deceptive Audio or Visual Media.” 2019. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730.

Afchar, Darius, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2019. “MesoNet: A Compact Facial Video Forgery Detection Network.” In *10th IEEE International Workshop on Information Forensics and Security, WIFS 2018*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/WIFS.2018.8630761.

Aguirre, Elizabeth, Dominik Mahr, Dhruv Grewal, Ko de Ruyter, and Martin Wetzels. 2015. “Unraveling the Personalization Paradox: The Effect of Information Collection and Trust-Building Strategies on Online Advertisement Effectiveness.” *Journal of Retailing* 91 (1): 34–49. https://doi.org/10.1016/j.jretai.2014.09.005.

Ahmed, Osman. 2021. “Balenciaga’s Deep-Fake SS22 Fashion Show Review - i-D.” 2021. https://i-d.vice.com/en_uk/article/akg3bb/balenciaga-ss22-show-review.

Amy Watson. 2021. “Statistics & Facts about Fake News | Statista.” 2021. https://www.statista.com/topics/3251/fake-news/#dossierKeyfigures.

Bayar, Belhassen, and Matthew C Stamm. 2016. “A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer.” In *IH and MMSec 2016 - Proceedings of the 2016 ACM Information Hiding and Multimedia Security Workshop*, 5–10. https://doi.org/10.1145/2909827.2930786.

Berthon, Pierre R., and Leyland F. Pitt. 2018. “Brands, Truthiness and Post-Fact: Managing Brands in a Post-Rational World.” *Journal of Macromarketing* 38 (2): 218–27. https://doi.org/10.1177/0276146718755869.

Bevans, Rebecca. 2020. “Akaike Information Criterion | When & How to Use It.” Scribbr. 2020. https://www.scribbr.com/statistics/akaike-information-criterion/.

Boerman, Sophie C., Eva A. van Reijmersdal, and Peter C. Neijens. 2014. “Effects of Sponsorship Disclosure Timing on the Processing of Sponsored Content: A Study on the Effectiveness of European Disclosure Regulations.” *Psychology and Marketing* 31 (3): 214–24. https://doi.org/10.1002/mar.20688.

Borau, Sylvie, and Marcelo Vinhal Nepomuceno. 2019. “The Self-Deceived Consumer: Women’s Emotional and Attitudinal Reactions to the Airbrushed Thin Ideal in the Absence Versus Presence of Disclaimers.” *Journal of Business Ethics* 154 (2): 325–40. https://doi.org/10.1007/s10551-016-3413-2.

Brewster, Thomas. 2021. “Fraudsters Cloned Company Director’s Voice In $35 Million Bank Heist, Police Find.” Forbes. 2021. https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/.

Burchard, Hans Von Der. 2018. “Belgian Socialist Party Circulates ‘deep Fake’ Donald Trump Video.” *Politico*, 2018. https://www.politico.eu/article/spa-donald-trump-belgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/.

Cahlan, Sarah. 2020. “How Misinformation Helped Spark an Attempted Coup in Gabon - The Washington Post,” February 13, 2020. https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/.

Campbell, Colin, Kirk Plangger, Sean Sands, and Jan Kietzmann. 2021. “Preparing for an Era of Deepfakes and AI-Generated Ads: A Framework for Understanding Responses to Manipulated Advertising.” *Journal of Advertising*. https://doi.org/10.1080/00913367.2021.1909515.

Campbell, Margaret C. 1995. “When Attention-Getting Advertising Tactics Elicit Consumer Inferences of Manipulative Intent: The Importance of Balancing Benefits and Investments.” *Journal of Consumer Psychology* 4 (3): 225–54. https://doi.org/10.1207/s15327663jcp0403_02.

Campbell, Margaret C., and Amna Kirmani. 2000. “Consumers’ Use of Persuasion Knowledge: The Effects of Accessibility and Cognitive Capacity on Perceptions of an Influence Agent.” *Journal of Consumer Research* 27 (1): 69–83. https://doi.org/10.1086/314309.

CereProc. 2018. “JFK Unsilenced | CereProc Text-to-Speech.” 2018. https://www.cereproc.com/en/jfkunsilenced.
Chaban, Matt. 2020. “Can Blockchain Block Fake News and Deep Fakes?” *IBM*, 2020. https://www.ibm.com/blogs/industries/blockchain-protection-fake-news-deep-fakes-safe-press/.
Chan, Caroline, Shiry Ginosar, Tinghui Zhou, and Alexei Efros. 2018. “Everybody Dance Now.” *Proceedings of the IEEE International Conference on Computer Vision* 2019-October (August): 5932–41. https://doi.org/10.1109/ICCV.2019.00603.
Chesney, Robert, and Danielle Keats Citron. 2018. “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security.” *SSRN Electronic Journal*, July. https://doi.org/10.2139/ssrn.3213954.
Chilton, Louis. 2020. “The Irishman’s Fan-Made de-Aging Deepfake Descibed as ‘Mind-Blowingly Better’ than the Original by Viewers | The Independent | The Independent.” 2020. https://www.independent.co.uk/arts-entertainment/films/news/irishman-deaging-deepfake-cgi-netflix-deniro-pacino-scorsese-video-a9686926.html.
Chollet, François. 2017. “Xception: Deep Learning with Depthwise Separable Convolutions.” In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:1800–1807. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/CVPR.2017.195.
Choudhry, Abhinav, Jinda Han, Xiaoyu Xu, and Yun Huang. 2022. “‘I Felt a Little Crazy Following a “Doll.”’” *Proceedings of the ACM on Human-Computer Interaction* 6 (GROUP): 1–28. https://doi.org/10.1145/3492862.
Christopher, Nilesh. 2020. “Deepfakes by BJP in Indian Delhi Election Campaign - VICE.” 2020. https://www.vice.com/en/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp.
Ciechanowski, Leon, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. 2019. “In the Shades of the Uncanny Valley: An Experimental Study of Human–Chatbot Interaction.” *Future Generation Computer Systems* 92 (March): 539–48. https://doi.org/10.1016/j.future.2018.01.055.
Cole, Samantha. 2017. “AI-Assisted Fake Porn Is Here and We’re All Fucked - Motherboard.” *Motherboard*, 1–12. https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn.
———. 2018. “Reddit Just Shut Down the Deepfakes Subreddit.” Motherboard. 2018. https://www.vice.com/en/article/neqb98/reddit-shuts-down-deepfakes.
Cossío-Silva, Francisco José, María Ángeles Revilla-Camacho, Manuela Vega-Vázquez, and Beatriz Palacios-Florencio. 2016. “Value Co-Creation and Customer Loyalty.” *Journal of Business Research* 69 (5): 1621–25. https://doi.org/10.1016/j.jbusres.2015.10.028.
Cozzolino, Davide, Giovanni Poggi, and Luisa Verdoliva. 2017. “Recasting Residual-Based Local Descriptors as Convolutional Neural Networks: An Application to Image Forgery Detection.” In *IH and MMSec 2017 - Proceedings of the 2017 ACM Workshop on Information Hiding and Multimedia Security*, 159–64. Association for Computing Machinery, Inc. https://doi.org/10.1145/3082031.3083247.
Csiszat, Gergo, Peter Varga Szilagyi, and Milievic Renata. 2022. “How Metaverse Will Shape the Future of Influencer Marketing.” 2022. https://influencermarketinghub.com/metaverse-influencer-marketing/.
Cury, Cecilia. 2021. “Spotify Wrapped: Lessons About Successful Marketing Strategies.” 2021. https://rockcontent.com/blog/spotify-wrapped-marketing-strategies/.
D’Angelo, Emanuele. 2021. “Can Deepfake Change FIFA’s Aesthetics Forever?” 2021. https://www.nssmag.com/en/sports/28215/deepfake-estetica-fifa.
Dahlén, Micael, Anton Granlund, and Mikael Grenros. 2009. “The Consumer-Perceived Value of Non-Traditional Media: Effects of Brand Reputation, Appropriateness and Expense.” *Journal of Consumer Marketing* 26 (3): 155–63. https://doi.org/10.1108/07363760910954091.
Darke, Peter R., and Robin J.B. Ritchie. 2007. “The Defensive Consumer: Advertising Deception, Defensive Processing, and Distrust.” *Journal of Marketing Research* 44 (1): 114–27. https://doi.org/10.1509/jmkr.44.1.114.
DARPA. 2021. “DARPA Announces Research Teams Selected to Semantic Forensics Program.” 2021. https://www.darpa.mil/news-events/2021-03-02.

Dave Gershgorn. 2020. “Snap Is Monetizing Deepfakes | OneZero.” 2020. https://onezero.medium.com/we-feared-deepfakes-then-tech-monetized-them-736746626d9c.
“Deepfakes/Faceswap: Deepfakes Software For All.” 2020. Github.Com. 2020. https://github.com/deepfakes/faceswap.
Dobber, Tom, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes de Vreese. 2021. “Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?” *International Journal of Press/Politics* 26 (1): 69–91. https://doi.org/10.1177/1940161220944364.
Doffman, Zak. 2019a. “Chinese {Deepfake} {App} {ZAO} {Goes} {Viral}, {Privacy} {Of} {Millions} ‘{At} {Risk}.’” *Forbes*. https://www.forbes.com/sites/zakdoffman/2019/09/02/chinese-best-ever-deepfake-app-zao-sparks-huge-faceapp-like-privacy-storm/#7d7775d28470.
———. 2019b. “Chinese Deepfake App ZAO Confirms Privacy Issue For Millions Of Users.” *Forbes*, 2019. https://www.forbes.com/sites/zakdoffman/2019/09/03/chinese-viral-app-zao-forced-to-backtrack-after-wechat-ban-and-privacy-backlash/.
Drescher, Daniel. 2017. *Blockchain Basics: A Non-Technical Introduction in 25 Steps*. *Blockchain Basics: A Non-Technical Introduction in 25 Steps*. Apress Media LLC. https://doi.org/10.1007/978-1-4842-2604-9.
Ducoffe, Robert H. 1995. “How Consumers Assess the Value of Advertising.” *Journal of Current Issues and Research in Advertising* 17 (1): 1–18. https://doi.org/10.1080/10641734.1995.10505022.
Dufour Nick, and Gully Andrew. 2019. “Google AI Blog: Contributing Data to Deepfake Detection Research.” *Google AI Blog*, 2019–21. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.
EUR-Lex. 2006. “EUR-Lex - 32006L0114 - EN - EUR-Lex.” 2006. https://eur-lex.europa.eu/eli/dir/2006/114/oj.
Fallis, Don. 2021. “The Epistemic Threat of Deepfakes.” *Philosophy and Technology* 34 (4): 623–43. https://doi.org/10.1007/s13347-020-00419-2.
Fowler, Geoffrey. 2019. “You Downloaded FaceApp. Here’s What You’ve Just Done to Your Privacy.” *The Washington Post*. https://www.washingtonpost.com/technology/2019/07/17/you-downloaded-faceapp-heres-what-youve-just-done-your-privacy/.
Friestad, Marian, and Peter Wright. 1994. “The Persuasion Knowledge Model: How People Cope with Persuasion Attempts.” *Journal of Consumer Research* 21 (1): 1. https://doi.org/10.1086/209380.
Gardiner, Nicholas. 2020. “Facial Re-Enactment, Speech Synthesis and the Rise of the Deepfake.” *The Grants Register 2021*, 344–45. https://ro.ecu.edu.au/theses_hons/1530.
Griskevicius, Vladas, Michelle N. Shiota, and Samantha L. Neufeld. 2010. “Influence of Different Positive Emotions on Persuasion Processing: A Functional Evolutionary Approach.” *Emotion* 10 (2): 190–206. https://doi.org/10.1037/a0018421.
Groh, Matthew, 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): e2110013119. https://doi.org/10.1073/pnas.2110013119.
Ham, Chang Dae. 2017. “Exploring How Consumers Cope with Online Behavioral Advertising.” *International Journal of Advertising* 36 (4): 632–58. https://doi.org/10.1080/02650487.2016.1239878.
Hancock, Jeffrey T., and Jeremy N. Bailenson. 2021. “The Social Impact of Deepfakes.” *Cyberpsychology, Behavior, and Social Networking* 24 (3): 149–52. https://doi.org/10.1089/CYBER.2021.29208.JTH.
Hastak, Manoj, and Michael B. Mazis. 2011. “Deception by Implication: A Typology of Truthful but Misleading Advertising and Labeling Claims.” *Journal of Public Policy and Marketing* 30 (2): 157–67. https://doi.org/10.1509/jppm.30.2.157.
Haysom, Sam. 2018. “People Are Using Face-Swapping Tech to Add Nicolas Cage to Random Movies and What Is 2018.” 2018. https://mashable.com/article/nicolas-cage-face-swapping-deepfakes.
He, Laura, Jack Guy, and Serenitie Wang. 2019. “Zao ‘deepfake’ App Backpedals after Privacy Backlash - CNN.” 2019. https://edition.cnn.com/2019/09/03/tech/zao-app-deepfake-scli-intl/index.html.

Heaven, Will Douglas. 2020. “The Year Deepfakes Went Mainstream | MIT Technology Review.” https://www.technologyreview.com/2020/12/24/1015380/best-ai-deepfakes-of-2020.
Helms, Karla Jo, and Daniel Mutter. 2020. “Deepfakes Could Mean Deep Losses for Businesses If Not Prepared.” 2020. https://www.prweb.com/releases/deepfakes_could_mean_deep_losses_for_businesses_if_not_prepared/prweb17431805.htm.
Hemlin Billström, Adam, and Fabian Huss. 2017. “Video Integrity through Blockchain Technology.” http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1129877&dswid=-8341.
Hitti, Natashah. 2020. “Samsung’s Artificial Neon Humans Are ‘a New Kind of Life.’” 2020. https://www.dezeen.com/2020/01/15/samsung-neon-star-labs-artificial-humans/.
Holubowicz, Gerald. 2020. “Le Deepfake de Christophe Guilhou, Embarrasse Au Cameroun - Journalism.Design.” 2020. https://journalism.design/le-deepfake-de-christophe-guilhou-embarrasse-au-cameroun/.
Hovland, Carl I., and Walter Weiss. 1951. “The Influence of Source Credibility on Communication Effectiveness.” *Public Opinion Quarterly* 15 (4): 635–50. https://doi.org/10.1086/266350.
Hoy, Mariea Grubbs, and Michael J. Stankey. 1993. “Structural Characteristics of Televised Advertising Disclosures: A Comparison with the Ftc Clear and Conspicuous Standard.” *Journal of Advertising* 22 (2): 47–58. https://doi.org/10.1080/00913367.1993.10673403.
Hyperledger. 2020. “A Blockchain Platform for the Enterprise — Hyperledger-Fabricdocs Master Documentation.” 2020. https://hyperledger-fabric.readthedocs.io/en/release-2.2/.
Infinitizer. 2018. “Infinitizer - Case Studies - Zalando.” 2018. https://www.infinitizer.de/en/case-studies/zalando-cara-calavigne/.
Jaiman, Ashish. 2020. “Positive Use Cases of Synthetic Media (Aka Deepfakes) | by Ashish Jaiman | Towards Data Science.” 2020. https://towardsdatascience.com/positive-use-cases-of-deepfakes-49f510056387.
Jesse Damiani. 2019. “A Voice Deepfake Was Used To Scam A CEO Out Of $243,000.” Forbes. 2019. https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/.
Jia, Ye, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2021. “Translatotron 2: Robust Direct Speech-to-Speech Translation.” https://google-research.github.io/lingvo-lab/translatotron2/.
Jia, Ye, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, et al. 2018. “Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis.” In *Advances in Neural Information Processing Systems*, 2018-Decem:4480–90. https://google.github.io/tacotron/publications/speaker_adaptation.
Johannes Tammekänd, John Thomas, and Kristjan Peterson. 2020. “Deepfakes 2020: The Tipping Point.” 2020. https://thesentinel.ai/media/Deepfakes 2020: The Tipping Point, Sentinel.pdf.
Kahn, Jeremy. 2021. “Forget Disinformation. Deepfakes Are Coming for Hollywood and Madison Avenue | Fortune.” 2021. https://fortune.com/2021/06/22/deepfakes-tom-cruise-chris-ume-metaphysic-hollywood-madison-avenue-eye-on-ai/.
Kelion, Leo. 2020. “Deepfake Detection Tool Unveiled by Microsoft - BBC News.” *Bbc*. https://www.bbc.com/news/technology-53984114.
Keltner, Dacher, and Jonathan Haidt. 1999. “Social Functions of Emotions at Four Levels of Analysis.” *Cognition and Emotion* 13 (5): 505–21. https://doi.org/10.1080/026999399379168.
Kietzmann, Jan, Linda W. Lee, Ian P. McCarthy, and Tim C. Kietzmann. 2020. “Deepfakes: Trick or Treat?” *Business Horizons*. Elsevier. https://doi.org/10.1016/j.bushor.2019.11.006.
Kietzmann, Jan, Adam J. Mills, and Kirk Plangger. 2021. “Deepfakes: Perspectives on the Future ‘Reality’ of Advertising and Branding.” *International Journal of Advertising* 40 (3): 473–85. https://doi.org/10.1080/02650487.2020.1834211.
Kleemans, Mariska, Serena Daalmans, Ilana Carbaat, and Doeschka Anschütz. 2018. “Picture Perfect: The Direct Effect of Manipulated Instagram Photos on Body Image in Adolescent Girls.” *Media Psychology* 21 (1): 93–110. https://doi.org/10.1080/15213269.2016.1257392.
Kline, Paul. 2000. “The Handbook of Psychological Testing,” 744.
Koh, Yoree, and Georgia Wells. 2018. “The Making of a Computer-Generated Influencer - WSJ.” 2018. https://www.wsj.com/articles/the-making-of-a-computer-generated-influencer-

11544702401.
Koppen, Camille, and Charles Spence. 2007. “Seeing the Light: Exploring the Colavita Visual Dominance Effect.” *Experimental Brain Research* 180 (4): 737–54. https://doi.org/10.1007/s00221-007-0894-3.
Korshunova, Iryna, Wenzhe Shi, Joni Dambre, and Lucas Theis. 2017. “Fast Face-Swap Using Convolutional Neural Networks.” In *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:3697–3705. https://doi.org/10.1109/ICCV.2017.397.
Kowalski, Marek. 2018. “GitHub - MarekKowalski/FaceSwap: 3D Face Swapping Implemented in Python.” https://github.com/MarekKowalski/FaceSwap/.
Krawitz, Marilyn. 2014. “Beauty Is Only Photoshop Deep: Legislating Models’ BMIs and Photoshopping Images.” *Journal of Law and Medicine* 21 (4): 859–74. https://www.researchgate.net/publication/264463496_Beauty_is_only_photoshop_deep_legislating_models’_BMIs_and_photoshopping_images.
KRO-NCRV. 2020. “Deepfake Therapy.” https://www.2doc.nl/documentaires/series/2doc/kort/2020/deepfake-therapy.html.
Kulp, Patrick. 2021. “Lay’s Lets Fans Customize a Deepfaked Message From Messi.” 2021. https://www.adweek.com/commerce/lays-lets-fans-send-personalized-video-messages-from-lionel-messi-with-deepfake-tech/.
Kwok, Andrei O.J., and Sharon G.M. Koh. 2021. “Deepfake: A Social Construction of Technology Perspective.” *Current Issues in Tourism*. Routledge. https://doi.org/10.1080/13683500.2020.1738357.
Lange, Fredrik, Sara Rosengren, and Angelica Blom. 2016. “Store-Window Creativity’s Impact on Shopper Behavior.” *Journal of Business Research* 69 (3): 1014–21. https://doi.org/10.1016/j.jbusres.2015.08.013.
Lee, Suk-hwan. 2021. “Blockchain Technology for Combating Deepfake and Protect Video / Image Integrity” 24 (8): 1044–58. https://www.koreascience.or.kr/article/JAKO202125761199587.pdf.
Leung, Gabrielle. 2020. “Spotify The Weeknd ‘Alone with Me’ AI Experience | HYPEBEAST.” 2020. https://hypebeast.com/2020/8/spotify-the-weeknd-personalized-artificial-intelligence-experience.
Little, Katie. 2014. “Why Restaurant Meals Don’t Look like the Ads.” CNBC. 2014. https://www.cnbc.com/2014/05/19/why-restaurant-meals-dont-look-like-the-ads.html.
Lomtadze, Ana. 2019. “The WSJ on DeepFakes: ‘It’s a Cat & Mouse Game’ | by Ana Lomtadze | Global Editors Network | Medium.” 2019. https://medium.com/global-editors-network/wsj-on-deepfakes-its-a-cat-mouse-game-212c3c7c6a4.
Mack, David. 2018. “This PSA About Fake News From Barack Obama Is Not What It Appears.” BuzzFeed News. 2018. https://www.buzzfeednews.com/article/davidmack/obama-fake-news-jordan-peele-psa-video-buzzfeed.
MacKenzie, Scott B., Richard J. Lutz, and George E. Belch. 1986. “The Role of Attitude toward the Ad as a Mediator of Advertising Effectiveness: A Test of Competing Explanations.” *Journal of Marketing Research* 23 (2): 130. https://doi.org/10.2307/3151660.
Melville, Kirsti. 2019. “The Insidious Rise of Deepfake Porn Videos — and One Woman Who Won’t Be Silenced.” ABC News. 2019. https://www.abc.net.au/news/2019-08-30/deepfake-revenge-porn-noelle-martin-story-of-image-based-abuse/11437774.
Mihailova, Mihaela. 2021. “To Dally with Dalí: Deepfake (Inter)Faces in the Art Museum.” *Convergence* 27 (4): 882–98. https://doi.org/10.1177/13548565211029401.
Mirsky, Yisroel, and Wenke Lee. 2021. “The Creation and Detection of Deepfakes.” *ACM Computing Surveys (CSUR)* 54 (1). https://doi.org/10.1145/3425780.
Mitchell, Amy, Jeffrey Gottfried, Galen Stocking, Mason Walker, and Sophia Fedeli. 2019. “Many Americans Say Made-up News Is a Critical Problem That Needs to Be Fixed.” *Pew Research Center* 5: 1–74. https://www.pewresearch.org/journalism/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/.
Modig, Erik, Micael Dahlén, and Jonas Colliander. 2014. “Consumer-Perceived Signals of ‘creative’ versus ‘Efficient’ Advertising: Investigating the Roles of Expense and Effort.” *International Journal of Advertising* 33 (1): 137–54. https://doi.org/10.2501/IJA-33-1-137-154.
Mohanta, Bhabendu Kumar, Soumyashree S. Panda, and Debasish Jena. 2018. “An Overview of

Smart Contract and Use Cases in Blockchain Technology." *2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018*, October. https://doi.org/10.1109/ICCCNT.2018.8494045.

Moon Jae-in. 2022. "Is South Korean the World's First Official Deepfake Candidate? Meet 'AI Yoon', the Salty AI Making the 2022 Presidential Election Interesting | South China Morning Post," February 14, 2022. https://www.scmp.com/news/asia/east-asia/article/3166928/south-korean-worlds-first-official-deepfake-candidate-meet-ai.

Mukherjee, Amaradri, Ronn J. Smith, and Anna M. Turri. 2018. "The Smartness Paradox: The Moderating Effect of Brand Quality Reputation on Consumers' Reactions to RFID-Based Smart Fitting Rooms." *Journal of Business Research* 92: 290–99. https://doi.org/10.1016/j.jbusres.2018.07.057.

Mulcahy, Rory, Kate Letheren, Ryan McAndrew, Charmaine Glavas, and Rebekah Russell-Bennett. 2019. "Are Households Ready to Engage with Smart Home Technology?" *Journal of Marketing Management* 35 (15–16): 1370–1400. https://doi.org/10.1080/0267257X.2019.1680568.

Murphy, Gillian, and Emma Flynn. 2021. "Deepfake False Memories." *Memory*. https://doi.org/10.1080/09658211.2021.1919715.

Panetta, Kasey. 2019. "Gartner Top Strategic Predictions for 2020 and Beyond." *Gartner*, no. October 2019: 1–13. https://www.gartner.com/smarterwithgartner/gartner-top-strategic-predictions-for-2020-and-beyond.

Paul, Kari. 2021. "Facebook Announces Name Change to Meta in Rebranding Effort." The Guardian. 2021. https://www.theguardian.com/technology/2021/oct/28/facebook-name-change-rebrand-meta.

Petty, Ross D., and J. Craig Andrews. 2008. "Covert Marketing Unmasked: A Legal and Regulatory Guide for Practices That Mask Marketing Messages." *Journal of Public Policy and Marketing*. SAGE PublicationsSage CA: Los Angeles, CA. https://doi.org/10.1509/jppm.27.1.7.

Plangger, Kirk, and Richard T. Watson. 2015. "Balancing Customer Privacy, Secrets, and Surveillance: Insights and Management." *Business Horizons* 58 (6): 625–33. https://doi.org/10.1016/j.bushor.2015.06.006.

Poon, J. 2021. *GANs Gone Wild: Public Perceptions of Deepfake Technologies on YouTube*. https://dspace.cuni.cz/handle/20.500.11956/150489%0Ahttps://dspace.cuni.cz/bitstream/handle/20.500.11956/150489/120397805.pdf?sequence=1.

Posner, Michael I., Mary J. Nissen, and Raymond M. Klein. 1976. "Visual Dominance: An Information-Processing Account of Its Origins and Significance." *Psychological Review* 83 (2): 157–71. https://doi.org/10.1037/0033-295X.83.2.157.

Powers, Katie. 2019. "Virtual Influencers Are Becoming More Real—Here's Why Brands Should Be Cautious." 2019. https://www.ama.org/marketing-news/virtual-influencers-are-becoming-more-real-heres-why-brands-should-be-cautious/.

Prajwal, K R, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C V Jawahar. 2020. "A Lip Sync Expert Is All You Need for Speech to Lip Generation in the Wild." In *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, 484–92. Association for Computing Machinery, Inc. https://doi.org/10.1145/3394171.3413532.

Rahmouni, Nicolas, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2017. "Distinguishing Computer Graphics from Natural Images Using Convolution Neural Networks." In *2017 IEEE Workshop on Information Forensics and Security, WIFS 2017*, 2018-Janua:1–6. https://doi.org/10.1109/WIFS.2017.8267647.

Rossler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. 2019. "FaceForensics++: Learning to Detect Manipulated Facial Images." In *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:1–11. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICCV.2019.00009.

Rothkopf, Joshua. 2020. "'Welcome to Chechnya' Brings Deepfake Technology to Documentaries - The New York Times." 2020. https://www.nytimes.com/2020/07/01/movies/deepfakes-documentary-welcome-to-chechnya.html.

Sayler, Kelley M., and Laurie A. Harris. 2019. "Deep Fakes and National Security (CRS Report No.

IF11333).” *Congressional Research Service*. https://apps.dtic.mil/sti/citations/AD1117081.
Shead, Sam. 2020. “Facebook to Ban ‘Deepfakes.’” *BBC News*, 2020. https://www.bbc.com/news/technology-51018758.
Shen, Jonathan, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, et al. 2018. “Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions.” In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018-April:4779–83. https://doi.org/10.1109/ICASSP.2018.8461368.
Shi, Tianyang, Yi Yuan, Changjie Fan, Zhengxia Zou, Zhenwei Shi, and Yong Liu. 2019. “Face-to-Parameter Translation for Game Character Auto-Creation.” In *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:161–70. https://doi.org/10.1109/ICCV.2019.00025.
Soh, Hyeonjin, Leonard N. Reid, and Karen Whitehill King. 2007. “Trust in Different Advertising Media.” *Journalism and Mass Communication Quarterly* 84 (3): 455–76. https://doi.org/10.1177/107769900708400304.
Statt, Nick. 2019. “China Makes It a Criminal Offense to Publish Deepfakes or Fake News without Disclosure - The Verge.” 2019. https://www.theverge.com/2019/11/29/20988363/china-deepfakes-ban-internet-rules-fake-news-disclosure-virtual-reality.
———. 2020. “TikTok Is Banning Deepfakes to Better Protect against Misinformation - The Verge.” 2020. https://www.theverge.com/2020/8/5/21354829/tiktok-deepfakes-ban-misinformation-us-2020-election-interference.
Thalen, Mikael. 2020. “Republican Candidate Calls George Floyd’s Death a Deepfake Hoax.” 2020. https://www.dailydot.com/debug/republican-candidate-george-floyd-conspiracy/?utm_medium=email&_hsmi=91059372&_hsenc=p2ANqtz--osmn58tvLTZD2M1jH4xfl6125WtgAjALoHdY2wSTXSGvDjWinhskL5wSPkbRnKW0qUxwR6TCGPqpPfeZnmTmS5gIX-qGRp3a2oONOTO5aiHDysuo&utm_content=91059372&utm_source=hs_email.
Thies, Justus, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. “Neural Voice Puppetry: Audio-Driven Facial Reenactment.” In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12361 LNCS:716–31. https://doi.org/10.1007/978-3-030-58517-4_42.
Thies, Justus, Michael Zollhöfer, and Matthias Nießner. 2019. “Deferred Neural Rendering: Image Synthesis Using Neural Textures.” *ACM Transactions on Graphics* 38 (4). https://doi.org/10.1145/3306346.3323035.
Thies, Justus, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2019. “Face2Face: Real-Time Face Capture and Reenactment of RGB Videos.” *Communications of the ACM* 62 (1): 96–104. https://doi.org/10.1145/3292039.
Ume, Chris. 2021. “Chris Ume on Instagram: ‘The Future of Gaming Is Here. This Is a Glimpse of What It “Could” Look like. Featuring #fifa21 Using a.i. Deepfake Technology.….’” 2021. https://www.instagram.com/p/CWV5NZGAk-r/.
Vega, Guillermo. 2021. “Una Campaña de Cruzcampo Con Un ‘Deepfake’ de Lola Flores Se Hace Viral | Tecnología | EL PAÍS.” EL PAÍS. 2021. https://elpais.com/tecnologia/2021-01-21/una-campana-con-un-deepfake-de-lola-flores-se-hace-viral.html.
Vincent, James. 2021. “Deepfake Dubs Could Help Translate Film and TV without Losing an Actor’s Original Performance - The Verge.” 2021. https://www.theverge.com/2021/5/18/22430340/deepfake-dubs-dubbing-film-tv-flawless-startup.
Westerlund, Mika. 2019. “The Emergence of Deepfake Technology: A Review.” *Technology Innovation Management Review*. https://doi.org/10.22215/TIMREVIEW/1282.
Whitaker, Bill. 2021. “Synthetic Media: How Deepfakes Could Soon Change Our World - 60 Minutes - CBS News.” 2021. https://www.cbsnews.com/news/deepfake-artificial-intelligence-60-minutes-2021-10-10/.
Whittaker, Lucas, Tim C. Kietzmann, Jan Kietzmann, and Amir Dabirian. 2020. “‘All around Me Are Synthetic Faces’: The Mad World of Ai-Generated Media.” *IT Professional* 22 (5): 90–99. https://doi.org/10.1109/MITP.2020.2985492.

Whittaker, Lucas, Kate Letheren, and Rory Mulcahy. 2021. “The Rise of Deepfakes: A Conceptual Framework and Research Agenda for Marketing.” *Australasian Marketing Journal* 29 (3): 204–14. https://doi.org/10.1177/1839334921999479.
Wiederhold, Brenda K. 2021. “Can Deepfakes Improve Therapy?” *Cyberpsychology, Behavior, and Social Networking*. Mary Ann Liebert Inc. https://doi.org/10.1089/cyber.2021.29209.editorial.
Williams, Robert. 2020. “Doritos Lets Mobile Fans Create Super Bowl Dance Videos with AI App | Marketing Dive.” 2020. https://www.marketingdive.com/news/doritos-lets-mobile-fans-create-super-bowl-dance-videos-with-ai-app/571361/.
Wu, Fuzhong, Yueran Ma, and Zheng Zhang. 2021. “I Found a More Attractive Deepfaked Self: The Self-Enhancement Effect in Deepfake Video Exposure.” *Cyberpsychology, Behavior, and Social Networking* 24 (3): 173–81. https://doi.org/10.1089/cyber.2020.0173.
Xie, Kevin, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. 2021. “Physics-Based Human Motion Estimation and Synthesis from Videos,” September. https://arxiv.org/abs/2109.09913v1.
Yu, Peipeng, Zhihua Xia, Jianwei Fei, and Yujiang Lu. 2021. “A Survey on Deepfake Video Detection.” *IET Biometrics* 10 (6): 607–24. https://doi.org/10.1049/bme2.12031.

Summary

In 2017, on Reddit, an unidentified user under the name of "deepfakes" uploaded some pornographic videos claiming they belonged to famous actresses such as Gal Gadot, Scarlett Johansson, Taylor Swift, Maisie Williams. The videos got labeled as fake but they were using a surprisingly accurate technology of face-swapping based on deep learning. Deepfakes started appearing in other Reddit r/deepfakes forums and got quickly noticed by news media like Vice that reported them (Cole 2017) (Cole 2018). The technology used open-source tools available from big tech companies like Nvidia and Google that made everything accessible, free of charge, and available at any time with little effort and knowledge. From that moment, deepfakes became available to anyone (Gardiner 2020). Reddit soon took out the videos, banned the user, and deleted r/deepfakes subreddit but it was too late to stop their spread.
The technology is evolving at a huge pace and it progressively requires less data and processing time to create one. In 2018 the app *FakeApp* required some consistent amount of input data and time to create one. In 2019, Zao, the Chinese app, only requires users to take a series of selfies with specific facial expressions and places whoever's face into a famous movie scene for free and in less than a minute (Kietzmann et al. 2020). Just a year later in 2020, a new application called Reface by Neocortext inc. allowed users to do the same as these two previously mentioned apps, but only with one selfie and in roughly 30 seconds.
Despite the fictional content like the ones that the mentioned apps offer, the technology could be used to cause more dangerous consequences. With the aid of proper adjustment, it's easy to trick the public into believing a fake video is real. Jordan Peele, in 2018, made Obama say that Donald Trump was "a total and complete dipshit". The result was hyper-realistic and frightening (Mack 2018). Videos might be fake, but the threat is so real DARPA (Defense Advanced Research Projects Agency) started two new projects devoted to developing Deepfake detection systems: Media Forensics (MediFor) in 2019, and Semantic Forensics (SemaFor) in 2020 (Sayler and Harris 2019).
In 2020, an Indian politician Manoj Tiwari, Bharatiya Janata Party (BJP) leader, collaborated with a communication agency that used a deepfake to make the politician speak in 3 different languages, Hindi, English, and Haryanvi (an Indian dialect spoken by most of his electors). This was the first time a deepfake has been used for an official campaign (Christopher 2020). The most recent example comes from just a couple of days ago. Korean politician Yoon Suk-yeol's team developed a platform where Yoon's deepfake answers questions to their potential electors (Moon Jae-in 2022).
From 2019, thesnetinel.ai reports, deepfakes have been having a year-over-year growth of ~6820x reaching more than 100M+ of total deepfakes online in 2020 (Johannes Tammekänd, John Thomas, and Kristjan Peterson 2020). Big tech companies like Snapchat, Apple, and ByteDance have integrated the function into their platforms (Dave Gershgorn 2020). New companies are entering the market offering deepfake solutions to create new business opportunities. Deepfakes entered the mainstream (Heaven 2020).

Different types of deepfakes

Facial Re-enactment

Face2Face

Facial re-enactment is the alteration of a target subject's facial expression in a video based on a source subject input. Face2face's approach was the first one to only require monocular RGB input in real-time. The algorithm employs a dense photometric consistency measure to track the facial expressions of both the source and target videos. Reenactment is therefore accomplished through the rapid and effective transfer of deformation between source and target. The target sequence is used to obtain the mouth interior that best fits the re-targeted expression and warp it to produce an

exact fit. Finally, it re-renders the synthesized target face on top of the associated video stream, blending it in with the real-world illumination. With this technique, it is possible to control, in real-time, the facial expression of the target subject that will follow what the RGB source will do (Thies et al. 2019).

Neural Voice Puppetry

Neural Voice Puppetry works following the same concept of face re-enactment, but instead of using video as a source, it uses an audio track. Starting from an audio sequence, the algorithm creates a photo-realistic output video of a target person that is in sync with the audio from the source input.
A deep neural network using a latent 3D face model space drives this audio-driven facial reconstruction. The model learns temporal stability inherently through 3D representation and uses neural rendering to generate photo-realistic output frames. This method is capable of using both voices of any unknown source or synthetic voices with text-to-speech algorithms (Thies et al. 2020).

Face Swapping Deepfakes

This type of deepfake superimposes the face of a source subject to the one of a target subject. The algorithm analyzes both the target and the source face creating a latent space of similar features that are then used to perfectly swap the face of the source to the target in the most credible way. It's one of the most common uses of deepfakes (like the one used in the subreddit r/deepfakes) and it's the type of deepfake that it's mostly known to the public (Korshunova et al. 2017).

Full Body Deepfakes

This type of deepfake expands the concept of facial re-enactment (Thies et al. 2019) to a full-body level by transferring the body movements of a source subject to the one of a target subject. Starting from a video of a person dancing, the algorithm can transfer the performance to a new target just after a couple of minutes of movements. The method employs video-to-video translation, with pose serving as an intermediary representation. It extracts poses from the source subject and applies the learned pose-to-appearance mapping to generate the target subject to transfer the motion (Chan et al. 2018)

Audio deepfakes

The AI analyzes a source subject from samples, and it is then able to recreate the synthesized voice which is then possible to use in a text-to-speech system recreating the voice of the source subject in any aspect, tone included (Jia et al. 2018).

Concerns, threats, risks

Truthiness is the attribution of validity based on how something seems or feels, independent from and regardless of its truthfulness (Berthon and Pitt 2018). The Sleeper Effect, states that even if consumers are aware that they have been exposed to fake content, the persuasive effects of that content have a lasting influence on later perceptions (Hovland and Weiss 1951).
Liar's Dividend, for example, could impact the perception of reality, confusing people that might not be able to recognize what's real and what is fake anymore (Chesney and Citron 2018).
The visual dominance and importance in respect to other signals is well known since 1976 (Posner, Nissen, and Klein 1976) and a study by Koppen and Spence in 2007 shows how visual information is the one humans rely most on (Koppen and Spence 2007).
The *epistemic threat* of deepfakes (Fallis 2021) coined by Don Fallis, explains how videos in later years have become the "gold standard" of truth. The ability to easily manipulate them though could change this standard and deepfakes would prevent people to believe what they saw.
Deepfakes' hyper-realism could be the cause of serious consequences under many dimensions. Starting from its origin, porn deepfakes could cause serious harm to targeted individuals. Telegram

underground communities offer a paid service to its users to be able to create customized pornfakes. These videos could be used in ransomfake attacks and cyberbullying (Johannes Tammekänd, John Thomas, and Kristjan Peterson 2020).
Companies too can be under a “deepfake attack”. In 2019, using a deepfake voice clone, scammers impersonated high-position managers and were able to convince the impersonated’s colleagues to approve/make huge transfers. The amount of money transferred was huge, the first time €220k was transferred, the second time $35M (Jesse Damiani 2019)(Brewster 2021).
Deepfakes could be responsible for much larger consequences. Reports have shown how the trend of trust in news outlets is progressively going down, and the confidence in recognizing fake news too (Amy Watson 2021). Deepfakes could be a big player in drastically further dropping the trust of the public towards news organizations and helping the spreading of fake news. The domino effect deriving from this tendency is worse than imaginable. If used improperly, deepfakes could lead to mass manipulation, election interfering, political uncertainty, and they could potentially undermine democracy. Fake videos could be easily used before elections to skew voters' opinions (Westerlund 2019). If used in a specific delicate time or targeted to a particularly sensitive group, tailor-made fake videos could be the spark of some attacks or other serious consequences on national security (Dobber et al. 2021).

Deepfake countermeasures

Many actions from many actors are required to fight the risks associated with deepfakes. A proper information and sensibilization campaign are necessary to make the public aware and literate about this technology, as well as detection systems and authenticity proofs. DARPA (Defense Advanced Research Projects Agency) started two new projects devoted to Automated Multimodal Media Manipulation detection gathering top-notch teams by Accenture Federal Services (AFS), Google/Carahsoft, New York University (NYU), NVIDIA, and Systems & Technology Research (DARPA 2021). Alongside these projects, universities and researchers from all over the world are trying to develop the best detection algorithm. Big improvements were possible thanks to a big dataset created for this scope in the FaceForensic++ (Rossler et al. 2019) where different algorithms were tested and benchmarked. Other studies suggest considering humans as a *collective intelligence* in combination with a detection algorithm. The AI is devoted to image analysis, while humans focus on spotting contextual flaws (Groh et al. 2022). The only problem with detection algorithms, though, is the fact they could end up creating a mouse vs cat never-ending chase. The developments and improvements of one algorithm will serve as a tool to further improve the other one, and so on.
A different point of view in fighting deepfakes is the use of blockchain to authenticate real videos instead of spotting fake ones. With its tamperproof records, logs, and transactions, it may be used to establish the validity and originality of digital media in a decentralized, trusted, and secure manner. The start-up Safe.press uses IBM's hyperledger fabric solution to check sources and track authenticity. A green safe.press stamp is placed on each page where a member publishes a press release or article. The stamp acts as a digital seal of approval that is tied to a blockchain key. The same concept can be applied to videos authenticity (Chaban 2020).

Opportunites

Besides the analyzed potential risks of this technology, many opportunities are possible with the implementation and good use of deepfakes. If from one side it harms democracy, from the other it could be used to help human rights activists and journalists to stay anonymous in dictatorial and oppressive regimes. For citizen journalists and activists, using technology to report injustices on traditional or social media may be incredibly empowering. Personal digital avatars provide autonomy and can assist individuals in expanding their purpose, thoughts, and beliefs, as well as enabling self-expression (Chesney and Citron 2018). Education and medicine could benefit too. A combination of deepfakes and AR/VR technology could be used to augment education or help therapy sessions. Students could interact with a deepfaked historical character and directly ask

him history questions (Jaiman 2020). Therapists, instead, could recreate virtual ambiances to better treat post-traumatic stress disorder (PTSD) in exposure therapy or even in rehabilitation therapy (KRO-NCRV 2020)(Wiederhold 2021).
Voice cloning deepfakes are already been used to help people who lost their voice due to medical conditions get it back. A startup called Project Revoice tries to achieve this using audio-text-to-speech technology. VOCALiD uses voicebanks and proprietary voice blending technology to generate distinct vocal personalities for those who have speech and hearing problems.
The media and film industry is certainly the one that could benefit the most from them. A deepfake revolution could change the whole industry. From de-aging to resurrecting or just starring famous actors/actresses, this technology could cover it all. A deepfake of a famous character will be less expensive than flying him/her out, paying for hospitality, and having them act (Kahn 2021). Actors in the future might just have a dataset of their faces ready to share with companies who require to use their image. Filmmakers could replace lines without having to reshoot the whole scene. Instead of having a full set and crew to fix minor changes, a tiny team might do the same with a computer. Face re-enactment could be implemented in real-time translations or when having to translate movies. The interpreter's or the character's voice can be blended into the target subject's face that will mimic the interpreter's lips or the dubbed voice (Vincent 2021) (Prajwal et al. 2020).
Following the revolution would be for sure the gaming industry. A combination of FaceSwapping techniques (Kowalski 2018) pose estimation (Xie et al. 2021) and full-body deepfakes (Kowalski 2018) could allow us to perfectly transfer our face and personal types of body movements into a video game character. Applications could vary from deepfaking yourself into videogames to needing less time to create virtual characters (Shi et al. 2019).

Marketing big deal: Opportunities and risks

Deepfakes could help marketers create deeper bonds and higher engagement with their customers and have a much broader reach (Whittaker et al. 2020). Zalando already used deepfake technology in 2018 with the #whereveryouare campaign creating a personalized ad featuring Cara Delevigne for every little town present in Europe, generating a 79M in increased revenue (Infinitizer 2018). Synthesia partnered with ad agency Craftww for JustEat's commercial featuring Snoop Dogg. Another version of the ad was created using deepfake technology that changed the lyrics and lip movements without having to take extra shots (Whitaker 2021). This ease of creation could go beyond the imaginable with hyper-personalized ads. People could have their avatars and biometrical information used by companies to create specifically hyper-tailored ads based on the data available. Fashion industries could use this technology to allow users deepfake themselves and try garments on in a virtual dressing room, or simply by being the model in model-fits picture previews on their e-commerce platforms. More easily and less threatening, brands could create apps that use deepfake technologies allowing customers to deepfake themselves into an already made advertisement (Whittaker et al. 2020) (Kietzmann, Mills, and Plangger 2021). In this way, customers can become co-creators of value in the advertising, shifting from being just a passive audience. The user not only becomes part of the ad and feels more engaged, but could also actively share the hyper-personalized ad starring him/her/they to social media platforms, helping the company to reach more potential quality customers (Kietzmann, Mills, and Plangger 2021). A study from 2017 suggests how there's a positive correlation between increased personalization and ad's persuasiveness and perceived benefit (Ham 2017) which "could increase the desired (re)action of the consumer" (Kietzmann, Mills, and Plangger 2021).
Brands could also generate synthetic influencers or endorsers to perfectly represent their brand and aesthetic creating their narrative and life drama, all without the unforeseen dangers or concerns that a human endorser can introduce (Koh & Wells, 2018; Powers, 2019). An existing famous example is Lil Miquela, built using CGI. With 3.1 Million Followers Lil Miquela has endorsed brands like Samsung, Prada, Calvin Klein, and Mini (Powers 2019). One of the problems of creating digital avatars was the Uncanny Valley Effect which is an unsettling feeling of unfamiliarity individuals experience when viewing or interacting with robots that are nearly but not quite identical

to humans (Ciechanowski et al. 2019). Deepfakes could overcome it by creating a more authentic and realistic avatar. Facial expressions play an important role in social interaction coordination (Keltner and Haidt 1999) and brands could deploy them to deliver promotional messages, help the customer during the shopping experience, or even become a proper virtual assistant (Hitti 2020) (Whittaker, Letheren, and Mulcahy 2021). In the Metaverse, brands could have virtual synthetic generated influencers to represent their customers. It's theoretically possible that one day we would have a philosophical conversation with Apple or Coca-Cola's generated avatars. The marketing opportunities are infinite in this new world (Mulcahy et al. 2019).

The implementation of VR headsets and deepfakes could help brands in several ways. Travel agencies for example could create super-resolution images of the destination and offer customers to be deepfaked in the environment. VR headsets could be implemented to give a little taste of what they would look like in the destination. Enabling customers to create a realistic visualization of their consumption experience results in a higher purchase intention (Kwok and Koh 2021) (Mulcahy et al. 2019).

The possibilities are unlimited, but before jumping into this technology, managers should be aware of the risks deriving from it. Data usage and privacy is the first thing. Consumers feel vulnerable when they realize their personal information has been collected without specific consent for creating personalized ads (Aguirre et al. 2015). Companies should put more effort into building trust with their customers by clearly disclosing everything about the collection and use of their information (Mukherjee, Smith, and Turri 2018). Customers are usually faced with what's referred to as *adversing calculus*, which is the weight scale customers use between benefits and costs of a service. If they feel the costs are higher than the benefits they would react negatively to the brand. A solution could be showing the value gained by the customer if they share certain data, inducing the customer to make the final choice (Plangger and Watson 2015). Something else to take into consideration that would need further study is the role of ad falsity when using deepfakes in marketing. Consumers are well known to negatively react to false information in advertisements in a sort of self-protecting mechanism (Friestad and Wright 1994). The mentioned study from Campbell et al. suggests and links other studies on fakery in general that could be easily applied to how consumers would perceive deepfakes in ads (C. Campbell et al. 2021). It's still not clear, though, how this technology would impact. What is most probably sure though, is how consumers weigh the perceived cost of an advertisement against the cost imposed on them in terms of time or effort (M. C. Campbell 1995). They react badly when they consider an ad to impose a comparatively high cost on them relative to the advertiser's costs. This could happen if they realize the cost of producing a deepfake ad is too little.

One last thing for companies to consider and anticipate is the fight against deepfakes by tech companies and governments. Facebook and Tiktok already banned deepfakes from their platforms (Shead 2020) (Statt 2020). Google and Microsoft are known to be working towards deepfake detection systems (Kelion 2020)(Dufour Nick and Gully Andrew 2019). China in 2019 made deepfakes without a disclosure a criminal offense just a month after California did a similar thing banning them during the election season (Statt 2019)(AB-730 2019). it's plausible more and more countries and governments will join the fight against them. Companies need to be aware of that before jumping too soon into investing an inconsiderate amount of money into this new thing which could be, eventually, completely banned from the mainstream. What could be more probable, though, is that it will be mandatory to disclaim its use.

Research question

It is reasonable to assume the grey area deepfakes are "living in" will soon become a more defined and colorful one. Some countries are already taking care of the issue and started creating laws that ban or regulate more in deep the use of it. This study wants to help understand how disclosure and its timing could impact consumers' attitude towards advertisement that uses deepfakes. In 2014, a paper revealed how disclosure timing impacts the persuasion of an. When the disclosure is placed before, or during the promotion, it negatively impacts the persuasion effect because the consumer

processes the information differently and becomes more critical. Creating a stronger resistance to the information conveyed (Boerman, van Reijmersdal, and Neijens 2014).
This research also wants to analyze if the term used for defining the technology could moderate the effect of disclosure timing on an ad's effect on a consumer's perception. “Deepfake” is a term that has been widely used in the colloquial world and by news media to warn about all the risks we analyzed. It’s possible the word is mentally linked to bad emotions and triggers a negative effect on consumers. “Synthetic media” and “deepfake” are the categorical variable that is going to be used as a moderator effect on disclosure timing which will be on three levels, before, during, and after.

The three research questions are:

Q1: Does deepfake disclosure affect consumers’ ad perception?

Q2: How does deepfake disclosure timing affect consumers’ ad perception?

Q3: Does using “synthetic media” instead of “deepfake” when disclosing moderate the effect of it?

The hypothesizes are:

H0: Ad perception is not affected by disclosure nor the use of different words.

H1: Disclosure negatively affects consumers’ ad perception

H2: If disclosure is positioned at the beginning, or during the ad, consumers will react more negatively than if placed at the end

H3: Using “Synthetic Ad” instead of “Deepfake” positively affects the attitude towards the ad.

Research methodology

The research began by manually scanning the web and gathering the needed media to be used for the stimuli. The videos needed to be ad deepfakes with a certain level of hyper-realism. 4 videos from 4 different brands were selected and assigned a code bit (Gillette = GIL, L’oreal = LOR, Jaguar = JAG, Truecaller = TCC). They have been then edited on Adobe Premiere Pro to create the disclosure messages for the 6 different scenarios made by combining the independent variable, categorical in three dimensions (Before = B_, During = D_, After = A_), with the moderator, categorical in two dimensions (Deepfake = DF_, Synthetic Media = SM_).

A code to be able to identify them later is created.

[Disclosure] + [Type] + [Brand]

For instance, the code for a scenario where the disclosure is shown during the Truecaller ad, using the term “Synthetic Media” would be:

[D_] + [SM_] + [TCC] = D_SM_TCC.

The survey was then built using logic conditions and randomizers to evenly distribute the stimuli across the respondents. The first part analyzed the perceived realism of the videos used in exposition 1, while the second part measured the attitude towards the ads shown. To measure the attitudes a 7-point Likert scale has been used and derived from the ADTRUST scale used by Soh, Reid, and King (Soh, Reid, and King 2007) as well as another question measuring to what extent they generally liked the ad is asked. The question asks “To what extent do you like or dislike this ad: (7-point Likert scale with like = 7; not like = 1)” (MacKenzie, Lutz, and Belch 1986).
In between exposition 1 and 2, an informative message explaining what Deepfake and Synthetic Media are is placed to fix the eventual knowledge asymmetry.

Data analysis

The survey received 377 responses. After data cleaning, only 300 were considered. The demographics analysis shows the participants are mostly located in Europe. The variable gender included – Female (61.3%), Male (35%), Prefer Not to Say (2%), Nonbinary / Third gender (1.7%). Most of the respondents' ages range between 25-34 (56%) and 18-24 (39%). Education-wise the most numerous have a bachelor's degree (39%), a Postgraduate Degree, or equivalent (36.7%). And finally, they are mostly Employed full time (44%) or are students (35.7%).
The pre-test confirmed the video chosen were perceived as real enough to easily trick the participants. Almost half of the participants wrongly confused the GIL deepfake video as real (46.8%) while a quarter was not able to tell if it was real or not (25.5%). The LOR video was the one that gave the best results. 81.2% of the time the deepfake video was confused as real, and only 9.4% correctly spotted its fakeness.

The main test's PCA analysis showed consistency with the items measured. The first 2 components explain 64.8% of the variance for exposition 1 and 65.9% for exposition 2. Cronbach alpha = .9 for exposition 1 and =.88 for exposition 2, confirming the consistency of the scale.
Different ANOVA models were tested. Akaike Information Criterion (AIC) test confirms for both exposition 1 and exposition 2 the Two-way ANOVA with variable BRAND as a blocking variable.

Two-way ANOVA + blocking variable: AD_TRUST = DISCLOSURE * TYPE + BRAND

The models fit the assumptions of homoscedasticity for both expositions and by looking at the plots and ANOVA results we can confirm DISCLOSURE is a significant independent variable with a p-value < .001 for exposition 1 and p-value < .05 for exposition 2.

Exposition 1 showed a significant difference between Before (B_) and During (D_) (p < .05), and between During (D_) and No disclosure (N_) (p < .5). Exposition 2 has slightly different results but the variable Before (B_) is still the most significant one which is significantly different compared to After (A_) with p < .05.

The model results suggest the variable TYPE has no moderating effect on AD_TRUST in both expositions.

Managerial implications and conclusions

Deepfakes are an emerging technology that can drastically reduce marketing costs and wide possibilities. AI-generated media can offer a new universe of solutions to marketers and managers. Disclosure regulations might come in the near future. Managers and marketers need to understand and study the best solution to avoid the risk to invest money and resources into something that might backfire or that it's simply not clear how it works on consumers and at the same time not being caught unprepared and losing a potential advantage.
Ad-hoc ads should be created to further investigate this topic precisely. A bigger sample size for a similar study and a smaller one with a neuromarketing approach using ECG and Eye-tracking could further help understand the reasons and brain processes behind the exposition of a deepfake to a human.
This study wants to provide a preliminary insight for managers and marketers on how to approach such a scenario by providing an analysis of the response consumers could have for different types of disclosure placements and terminology used. The results demonstrated that a disclosure message does affect the overall attitude towards the ad, but it also showed that, opposed to what was hypothesized, placing it before the ad impacts in a much smaller magnitude compared to placing it during, or after it. Lastly, the difference in the terminology used doesn't seem to play a significant role as expected.

[1] https://apps.apple.com/it/app/reface-cambia-faccia/id1488782587

[2] https://www.youtube.com/watch?v=cQ54GDm1eL0

[3] https://www.darpa.mil/

[4] https://www.instagram.com/p/ByaVigGFP2U/

[5] https://www.tiktok.com/@deeptomcruise

[6] https://www.vice.com/en/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp

[7] https://www.resemble.ai/

[8] https://www.descript.com

[9] https://www.vooruit.be/nl/

[10] https://www.facebook.com/Vlaamse.socialisten/videos/10155618434657151/

[11] https://www.france24.com/en/

[12] https://www.facebook.com/watch/?ref=external&v=324528215059254

[13] https://web.telegram.org/

[14] https://www.resemble.ai/

[15] https://www.descript.com

[16] https://investigatedeepfakefloyd.com/

[17] https://scholar.google.com/citations?user=WgyrxUAAAAAJ&hl=it

[18] https://www.youtube.com/channel/UCbfYPyITQ-7l4upoX8nvctg

[19] https://www.darpa.mil/

[20] https://docs.soliditylang.org/en/v0.8.11/

[21] https://safe.press/

[22] https://wallpaperaccess.com/full/3978280.jpg

[23] https://thedali.org/

[24] https://www.microsoft.com/it-it/hololens

[25] https://www.projectrevoice.org/

[26] https://vocalid.ai/

[27] https://www.flawlessai.com/product

[28] https://www.youtube.com/watch?v=QiiSAvKJIHo

[29] https://www.hbo.com/

[30] https://www.youtube.com/c/Shamook

[31] https://www.netflix.com/

[32] https://github.com/iperov/DeepFaceLab

[33] https://www.cruzcampo.es/

[34] https://en.wikipedia.org/wiki/Lola_Flores

[35] https://metaphysic.ai/

[36] https://www.instagram.com/

[37] https://www.ea.com/en-au/games/fifa/fifa-22

[38] https://www.ea.com/en-au

[39] http://www.neteasegames.com/

[40] https://www.zalando.it/

[41] https://www.artificialtalent.co/

[42] https://www.balenciaga.com

[43] https://www.youtube.com/c/NEONdotlife/videos

[44] https://synthesia.io/

[45] https://www.craftww.com/

[46] https://www.justeat.com/

[47] https://en.wikipedia.org/wiki/Snoop_Dogg

[48] https://www.menulog.com.au/

[49] https://www.lays.com/

[50] https://www.messimessages.com/

[51] https://www.doritos.com/

[52] https://getsway.app/index.html

[53] https://www.washingtonpost.com/

[54] https://www.forbes.com/

[55] https://www.youtube.com/channel/UC_aeAIzNuRgHDDAd8loQp_A

[56] https://www.youtube.com/channel/UCdGx1hd1W76RXeuKrGnC3Kg

[57] https://www.microsoft.com/it-it/microsoft-365/excel

[58] https://colab.research.google.com/

[59] https://github.com/iperov/DeepFaceLab