TALLINN UNIVERSITY OF TECHNOLOGY

School of Information Technologies

Dominika Helena Jantas 201673IVCM

# Strategic considerations to counter the cyber threat of malicious deepfakes in Greek society.

Master's thesis

Supervisor: Adrian Venables

PhD

Tallinn 2022

# Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Dominika Helena Jantas

# Abstract

The issue of deepfakes and their illicit use has recently become increasingly prominent. Developments in technology and Artificial Intelligence have caused a breakthrough in the manipulation of audio-visual content. Additionally, social media and the Internet have provided a fertile environment for malicious deepfakes to spread quickly. The technology for deepfake synthesis is becoming cheaper and more accessible. Malicious actors are utilising latest technological advances to increase the sophistication of their deception and cases where malicious deepfakes are being used against individuals or organisations are increasing. As a result, the concerns of governments, legislative bodies, private sector, and Internet users are growing, and the topic is a current area of discussion by cybersecurity experts in Greece. This study focuses on the current deepfakes landscape in Greece and examines in detail the current legislation to address the threat it presents. It also investigates the societal perceptions of the Greek online community and aims to measure their exposure to malicious deepfakes and the implications their cybersecurity. In addition, the study contributes to the research on deepfakes detection by presenting two novel detection methods implemented in Greece. The findings of this research contribute to a better understanding of this cyber threat. These results could be taken into consideration by policy and law makers to assess the level of the deepfake threat in the Greek cyber domain. They also offer a contribution to the development of a future deepfakes strategy and options for the prevention and mitigation of malicious deepfakes in the Greek online domain.

This thesis is written in English and is 89 pages long. It includes 9 chapters, 50 figures, and 77 tables. It also contains 4 Annexes.

# Annotatsioon

Viimastel aegadel on süvavõltsingute ja nende ebaseadusliku kasutamise probleem muutunud üha olulisemaks. Tehnoloogia ja tehisintellekti areng on põhjustanud läbimurde audiovisuaalse sisuga manipuleerimises. Lisaks on sotsiaalmeedia ja Internet loonud soodsa keskkonna pahatahtlike süvavõltsingute kiireks levimiseks. Süvavõltsingute sünteesi tehnoloogia muutub odavamaks ja kättesaadavamaks. Pahatahtlikud osalejad uurivad uusimaid tehnoloogia edusamme, et oma pettusi veelgi keerukamaks muuta, ning juhtumid, kus üksikisikute või organisatsioonide vastu kasutatakse pahatahtlikke süvavõltsinguid, sagenevad. Selle tulemusena kasvavad valitsuste, seadusandlike organite, erasektori ja Interneti-kasutajate mured ning see teema on Kreekas küberjulgeolekuekspertide praegune aruteluvaldkond. See uuring keskendub praegusele süvavõltsingute maastikule Kreekas ja uurib üksikasjalikult kehtivaid sügavvõltsinguid käsitlevaid õigusakte. Samuti uurib see Kreeka veebikogukonna ühiskondlikke arusaamu ja selle eesmärk on mõõta selle kokkupuudet pahatahtlike süvavõltsingutega ja hinnata süvavõltsingute tagajärgi kasutajate küberturvalisusele. Lisaks aitab uuring kaasa sügavate võltsingute tuvastamise uurimisele, tutvustades kahte Kreekas rakendatud uudset tuvastamismeetodit. Uuringu tulemused võivad aidata kaasa selle küberohu paremale mõistmisele. Poliitika- ja seadusandjad võiksid neid tulemusi arvesse võtta, et hinnata süvavõltsimise ohu taset Kreeka kübervaldkonnas. Samuti tõstavad nad esile tulevase süvavõltsingute strateegia koostist Kreekas ja annavad ülevaate pahatahtlike süvavõltsingute ennetamise ja leevendamise võimalustest Kreeka veebidomeenis.

See lõputöö on kirjutatud inglise keeles ja on 89 lehekülge pikk. See sisaldab 9 peatükki, 50 joonist ja 77 tabelit. Sellel on ka 4 lisa.

# Acknowledgements

First and foremost, I would like to express my special thanks of gratitude to my supervisor, Dr. Adrian Venables for the continuous support of my study and research, for his patience, motivation, enthusiasm, and immense knowledge. Without his superb skills, good cheer and mentorship, this thesis would not be written. Monday mornings and meetings have never been so enjoyable and looked forward by me.

I dedicate this thesis to my parents, Helena and Ryszard Jantas, whose encouragement and good advice were motivators for my studies. I thank them for the years of patience and unconditional support of every endeavour I had. Only they know the obstacles, and hours of effort I spent on my studies. They also enriched my character and actions with values of hard work, commitment to self-growth, kindness, and genuineness. I grabbed my joy and passion for perpetual learning and exploration from them. I hope they are proud of me.

This thesis is also dedicated to Kamila Jantas, my twin sister and partner in crime since 1996. She has helped me to be a more creative, honest, and open person. Her presence and emotional support during my life and the years of my master’s studies were irreplaceable.

Special thanks are dedicated to my older sister, Sara Jantas, whose sense of humour and great online company made my experience with thesis sweet and delighted. Her academic experience and critical point of view lifted the quality of the thesis.

Finally, I gratefully thank my dear friend George Kyrkos. He stood by my side, motivated me, and explored with me the world of data analytics. His teaching skills are impressive, and his continuous self-growth and integrity of character are admirable by me. I hope he enjoys reading this thesis.

*“Η παιδεία είναι πανηγύρι της ψυχής, γιατί σ’ αυτήν υπάρχουν πολλά θεάματα και ακούσματα της ψυχής.” – Sokrates*

# List of Abbreviations and terms

| | |
|---|---|
| A2V | Audio-to-Video |
| AI | Artificial Intelligence |
| AV | Audio-Visual |
| CEO | Chief Executive Officer |
| CPU | Central Processing Unit |
| Conv-LSTM | Convolutional Long Short-Term Memory |
| DCT | Discrete Cosine Transform |
| DNN | Deep Neural Network |
| EAR | Eye-Aspect Ratio |
| EBV | Eye Blinking Video |
| EM | Expectation Maximisation |
| ENISA | European Union Agency for Cybersecurity |
| EU | European Union |
| FFHQ | Flickr-Faces-HQ |
| FSGAN | FaceSwaping Generative Adversarial Network |
| GAN | Generative Adversarial Network |
| GAN DCT | Generative Adversarial Network Discrete Cosine Transform |
| GANSF | Generative Adversarial Network Specific Frequency |
| HMM | Hidden Markov Models |
| HOHA | Hollywood Human Actions |
| ICT | Information and Communication Technologies |
| IT | Information Technology |
| KNN | K-Nearest Neighbour |
| LDA | Linear Discriminant Analysis |
| LRCN | Long-term Recurrent Convolutional Network |
| LSTM | Long Short-Term Memory |
| ML | Machine Learning |
| MSU | Michigan State University |
| NCSI | National Cybersecurity Index |
| PPG | Photoplethysmography |

| | |
|---|---|
| RNN | Recurrent Neural Network |
| ROC | Receiver Operating Characteristic |
| ROI | Region of Interest |
| SPPS | Statistical Package for the Social Sciences |
| T2V-L | Text-To-Video with long statements |
| T2V-S | Text-To-Video with short statements |
| VDE | Video Dialogue Replacement |

# Table of contents

# List of figures

# List of tables

# 1. Introduction

Deepfakes are considered a significant threat for the 21$^{st}$ century Information Society with many commentators warning that the risk of malicious deepfakes is on the rise [1]. As such, the dangers that they pose are  now being considered by governments and law making bodies as well as  by social media and Internet users [2] [3] [4] [5] [6]. The issue of deepfakes is now being widely discussed by governments around the world [7] [8].

The use of deepfakes has migrated from the beneficial use cases of the entertainment and games industry, education, fashion, e-commerce, and technology to malicious and more deeply concerning activities [9]. These include cybercrime and fraud, identity theft and misinformation [10]. In an example of theft, an audio deepfake of a company's CEO was used to authorise the transfer of money to a criminal, resulting in significant financial loss [11]. This was not the first case of a voice deepfake used in cyber fraud [12].  Synthetically generated images of people have also been used to create fake social accounts. [13]. Two prominent examples were 'Maisy Kinsley' and 'Katie Jones'. Both were fake personas with profiles on LinkedIn and Twitter respectively, which were involved in espionage campaigns [13]. These cases illustrate the increasing range of attack vectors available to malicious actors in cyberspace. These tend to lean towards utilizing the latest technology advancements, such as deepfakes, to target individuals and companies.

The commodification of deepfakes technology has already been seen. The source code from a creator of deepfakes was uploaded on Github and made accessible to the online community [14]. Numerous code libraries were subsequently published to enhance the quality, efficiency, and usability of the code [14]. Software and tools to create deepfakes have also been distributed online. Service portals selling customised deepfakes and advertisements of individual creators in online marketplaces are multiplying [14]. The industry of deepfakes is flourishing and is becoming better and increasingly accessible to a wide audience, including malicious actors.

## Motivation

The creation of malicious deepfakes targeted specifically at Greek users is one of the threats being considered by the country's cybersecurity community [15]. Since 2019, Greece has ranked first among 160 countries in the National Cybersecurity Index (NCSI) and has continuously sought to ensure cyber safety and develop secure websites for its regular users [16]. The Greek National Cybersecurity Strategy is also considered as one of the most comprehensive across the EU [17]. The country is also currently 28th on Global Cybersecurity Index and 38th on the (Information and Communication Technologies (ICT) Development Index. All these show that the development and commitment towards cybersecurity have been strongly promoted in Greece [16]. Despite that, the cyber threat of deepfakes is yet to be addressed by research and law and policy makers in Greece.

## Research purpose

The purpose of this thesis is to provide an extensive analysis of the use of deepfakes in Greece. The elements that form this landscape are government legislation, societal perceptions and detection strategies that address the threat of deepfakes in Greece. Also, the aim is to propose a national strategy which could regulate the use of this technology in Greece.

## Research questions

This thesis contains 5 research questions. It is accepted that this is a large number for a thesis of this type. However, it is considered that to fully understand the deepfakes landscape in Greece all these issues need to be addressed.

**RQ1.** What is the current legislation in Greece on deepfakes?

**RQ2.** Can Greek Internet and social media users distinguish deepfake videos and images from real ones?

**RQ3.** What is the impact of malicious deep fakes on the Greek Internet and social media users, based on their perceptions and experiences in the cyber domain?

**RQ4.** What is the effectiveness of deepfakes detection methods used in Greece and their perceived contribution to the cybersecurity of the Greek online community?

**RQ5.** How can the issues related to the malicious use of deepfakes be addressed in a Greek deepfakes strategy?

## Scope and goal

The scope of the study is the threat from deepfakes in Greece. It focuses geographically only on Greece. The population analysed is the English-speaking component of Greek society and specifically social media and Internet users. The area of discussion is the perception of the Greek online community towards deepfakes and its impact on their cybersecurity.

The legislation that will be reviewed originates from Greece, but also from the European Union. Finally, the strategies addressing the problem of deepfakes in Greece focus on two interesting approaches introduced by Facebook and Ellinika Hoaxes, detailed in Annex D.

The key assumptions are:

- There is no legislation in Greece regulating deepfakes technology. In late 2021, the EU started drafting laws on Artificial Intelligence, which is an international first [18] .
- Greek social media and Internet users have high levels of detection accuracy. A survey from 2020 showed that 72.8% of Greeks participants were capable of distinguishing fake news [19].
- Previous research concluded that automatic detection tools are short lived and cannot be effective in long term due to continuous improvements of deepfakes synthesis [20] [21].Thus, the solution to the problem of deepfakes is more complex and should search beyond detection technologies. A mix of legislation, user education and policies from the government, and regulations from private sector have been implemented in the USA with great success [22]. Such a

comprehensive strategy can be introduced in Greece to address the issue of deepfakes.

## Novelty

The use of deepfake technology is growing rapidly online [23]. Despite its enhanced presence, there has not been any comprehensive study on the topic conducted in the case of Greece. This is even though the country is within the European Union and a key partner in social media and Facebook cybersecurity research, including deepfakes.

There are no studies which investigate the legislation implemented in Greece on deepfakes. Besides the legislation, several studies have investigated the perceptions of Greek users towards the social media platforms and Internet [24] [25] [26] [27]. However, there is no research on the deepfake detection accuracy of Greek users, their exposure to deepfakes online and their perceptions towards deepfakes, and their cybersecurity. To date, there has been no examination of deepfakes detection specific to Greece in the available literature. Previous research only addressed fake news detection [19]. This expanding analysis investigates the wider issues of deepfakes. This research looks to investigate any short terms trends or confirm the results of the previous analysis.

Thus, this study investigates the topic, by examining the deepfake technology landscape in Greece. It aims to provides a clearer picture of the perceptions towards malicious deepfakes and their implications on the level of cybersecurity of Greek users. It also gives an overview of the detection methods to protect users in Greece from malicious deepfakes. Finally, the thesis tries to shape a future Greek deepfakes strategy.

This thesis will contribute to the understanding of the deepfake landscape in Greece, which is a novel area of research. The results of this research will provide a deeper appreciation of the risk of deepfakes in Greece and the ability of the population to detect potentially malicious material. The findings of the research could be of use in contributing to future legislation on how to regulate the use of the deepfakes technology. With the use of malicious deepfakes and therefore the threat they pose increasing, this research is both timely and highly relevant.

## Chapter overview

Chapter 2 is dedicated to describing the methodology used to conduct the study. A Literature review is used to explore the current deepfakes legislation as well as the deepfakes detection methods implemented in Greece. A questionnaire is used to assess the threat of deepfakes on Greek social media users and discuss the implications on their cybersecurity.

Chapter 3 provides an overview of the relevant literature on deepfakes. Topics covered include the rise of deepfakes, typology and generation process, implications of deepfakes, technologies and human ability as deepfakes detection methods.

Another topic covered is the legislation in Greece and the European Union focusing on the threat of deepfakes. This analysis is expanded Annex A and it is acknowledged that the focus on legal issues could be the subject of a separate research. However, its inclusion is made to provide a comprehensive understanding of the subject and related issues. Chapter 4 and Annex C focus on the results of the interviews conducted within Greek social media and Internet users. These sections provide a clearer picture of the deepfakes state within the Greek society, based on its perceptions and experiences. The exposure rate of users to deepfakes, the proportion of malicious and non-malicious deepfakes on the Internet and social media are shown. Users are also tested to detect deepfake images and videos. Annex B contains the Survey that was used to determine how accurately the English speaking Greek online community could identify both image and video based deepfakes. The analysis of all results obtained is in Chapter 5. Chapter 6 serves as a section for discussion on the results and provides answers to the research questions. It also includes suggestions for a National Deepfake Strategy. These could provide valuable advice for policy and legislation makers in Greece to construct a comprehensive strategy on deepfakes. Chapter 7 presents areas for future research and Chapter 8 summarises the fundings of the research to clarify what was learned from the thesis. The topic of deepfakes detection methods used in Greece is contained in Annex D including a description of a detection algorithm used by Facebook. The approach used by Ellinika Hoaxes in relying on human detection abilities is also analysed in Annex C.

# 2. Methodology

The chapter describes the methodology used to investigate the deepfakes landscape in Greece and address the research questions. This methodology sought to answer the following research requestions of this thesis.

**RQ1.** What is the current legislation in Greece on deepfakes?

**RQ2.** Can Greek Internet and social media users tell apart deepfake videos and images from real ones?

**RQ3.** What is the impact of malicious deep fakes on the Greek Internet and social media users, based on their perceptions and experiences in the cyber domain?

**RQ4.** What is the effectiveness of the deepfakes detection methods used in Greece and their perceived contribution to the cybersecurity of the Greek online community?

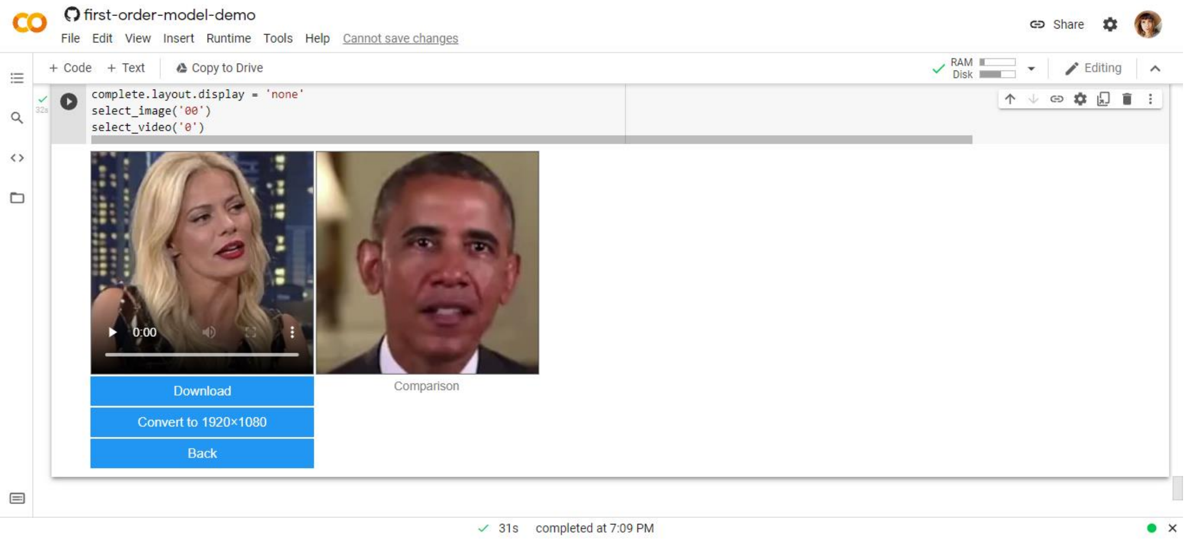
**RQ5.** How can the issues related to the malicious use of deepfakes be mitigated in a Greek deepfakes strategy?

The method used to address the first research topic is the literature review contained in Annex A, combined with a detailed examination of official governmental and legal documents of Greece. Specifically, the sources of this data originated from government publications and online papers published by Greek organisations such as LawSpt.gr, which explain legal matters.

The aim was to determine if the Greek legislative system foresees the needs for regulations on deepfakes technology. As Greece is a member of the European Union its legislation on deepfakes has been incorporated into the research process of the legislation applied in Greece. The legislation of the European Union was obtained from the EUR-Lex [91] and Europa [92] websites. The main benefit of this approach is that it provided a detailed collection of relative laws and regulations and gave an answer to whether Greece has deepfakes legislation.

The approach to the second and third research questions relied on data analysis. Specifically, social media and Internet users from Greece were asked to participate in an online questionnaire survey. The targeted audience included private individuals, regardless of gender, age, educational and professional background. The only condition was that they regularly used the Internet and social media and are living in Greece. The language of the target audience was limited to the English speaking community. To reach a larger audience and maximise the response rate, not only were social media platforms used but also email was utilised for this purpose. The response rate was circa 80%, with **150** invitations sent and **123** responses received. A pilot survey on 10 individuals was also used to test the proposed approach.

The questionnaire was divided into three sections. The first one presented a series of images or videos, some of which are deepfakes and others real images. The participant was asked to determine if the content shown is real or digitally manipulated. The deepfakes videos were created with the use of the First Order Model provided by Aliaksandr Siarohin and available at [93] and [94]. All steps of the deepfake creation process were performed by running the code cells and can be seen on the website https://colab.research.google.com/github/AliaksandrSiarohin/first-order-model/blob/master/demo.ipynb#scrollTo=Oxi6-riLOgnm [94]. The object of the deepfake was derived from a chosen source image. A source video was also selected to provide the motion for the source image in the final deepfake video. The deepfake was created within few minutes. The following Figure 1 shows the final step of the deepfake creation.



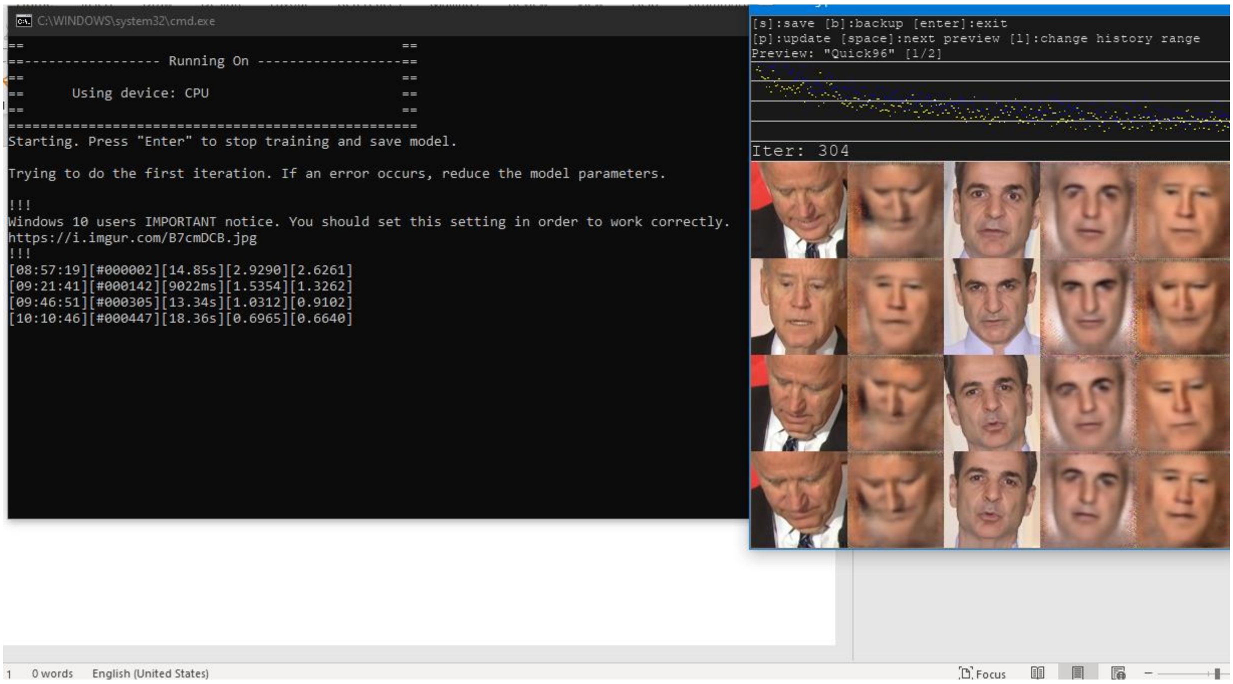
```
complete.layout.display = 'none'
select_image('00')
select_video('0')
```

Figure 1

. Deepfake created with the First Order Model

As for the deepfake images, one image was selected from the website https://www.whichfaceisreal.com/index.php. The rest of the deepfakes were downloaded from the Internet as genuine images and then manipulated in the Android application FaceApp [95]. The authentic images and video were chosen from free photos available online. The content aimed to be familiar to Greeks with the personalities in the images and videos being popular and recognisable. The main reason for this selection is that it better reflected the real-world online environment in which deepfakes present celebrities and politicians [96].

At this point, it is crucial to note that the initial goal was to create deepfake videos with the use of the *DeepFaceLab* [97]. The following Figure 2 depicts the process of AI training with the use of this software.



C:\WINDOWS\system32\cmd.exe

```
==                                                   ==
==------------------ Running On -------------------==
==                                                   ==
==      Using device: CPU                            ==
==                                                   ==
=====================================================
Starting. Press "Enter" to stop training and save model.

Trying to do the first iteration. If an error occurs, reduce the model parameters.

!!!
Windows 10 users IMPORTANT notice. You should set this setting in order to work correctly.
https://i.imgur.com/B7cmDCB.jpg
!!!
[08:57:19][#000002][14.85s][2.9290][2.6261]
[09:21:41][#000142][9022ms][1.5354][1.3262]
[09:46:51][#000305][13.34s][1.0312][0.9102]
[10:10:46][#000447][18.36s][0.6965][0.6640]
```

```
[s]:save [b]:backup [enter]:exit
[p]:update [space]:next preview [l]:change history range
Preview: "Quick96" [1/2]
Iter: 304
```

Figure 2. DeepFaceLab training process

The biggest bottleneck was the Central Processing Unit (CPU) usage. Specifically, the possessed CPU Intel (R) Core (TM) i5-3337U CPU @ 1.80GHz was overstrained with the DeepFaceLab tasks and the process of training finished at around 6000 iterations. Then, after approximately 10-11 hours, the machine crashed. Also, the use of CPU for AI training was time-consuming. The machine worked continuously for 4 days on frames and facets extractions from source and destination videos and 2 days for their training. It must be mentioned that the source and destination videos were aligning in terms of length, pixels and the angles and positions of the individuals presented.

The solution to the above issue was the *First Order Model*. The main benefit of this model was that the machine's CPU was not involved in the process of deepfake creation. The deepfake was created with simple code execution. Also, the process to create a deepfake took less than 5 minutes. The drawbacks of this method were the quality of the deepfake which was not as high as the publicly available deepfakes created with the use of DeepFaceLab [98] [99] [100]. At the same time, the voice could not be a feature of the deepfakes created with the First Order Model. The DeepFaceLab software maintained the voice of the source video in the destination video as well.

The second part of the questionnaire consisted of a series of multiple-choice questions. They aimed to investigate the habits, preferences, and behaviours of the Greek online community. Moreover, it measured the social media and Internet consumption of Greeks and the popularity of the social media platforms and Internet websites. Other observations derived were the frequency measurements of users' exposure to deepfakes online. Finally, the goal was to find out which website exposed Greek users to deepfakes at the highest level.

The third part of the questionnaire involved a series of questions or statements requiring a response using a 5-point Likert scale (from 1 as Strongly Disagree to 5 as Strongly Agree). The questions were designed to assess the participants' opinions on the use and impact of deepfakes on their cybersecurity and trust towards the virtual environment. Some questions aimed to explore the opinions of Greek users on the effectiveness and confidence towards the deepfakes technologies used by social media platforms, such as the Facebook Artificial Intelligence (AI) Michigan State University (MSU) tool. Besides this, the intent was to portray the level of their satisfaction towards the existing cybersecurity policy and efforts in Greece. They expressed their opinion on the Greek government and Ellinika Hoaxes regarding the mitigation or prevention of malicious deepfakes online.

Reliability analysis of the proposed instrument was performed with the use of Statistical Package for the Social Sciences (SPSS), while Cronbach's alpha (α) was leveraged as a measure of the internal consistency of the questionnaire conducted. According to Tavakol and Dennick, "Chronbach's alpha provides a measure of the internal consistency of a test or scale” [101]. Internal consistency is “the extent to which all the items in a test measure the same concept or construct and hence it is connected to the inter-relatedness of the items within the test” [101].

The research method combined both qualitative and quantitative research data within a single framework. Data collected from the Linkert scale ratings were quantitative [102]. The types of social media accounts that users have, and the categories of the websites accessed by the participants were examples of qualitative data.

The combination of both methods is extremely beneficial [103]. Qualitative research allows flexibility and documents human attitudes and routines, building patterns and trends related to the target group [103]. Moreover, it enables the different dimensions of the problem to be analysed [104]. In quantitative research, the collected data can be measured [104]. It systematically incorporates structured procedures and formal tools for data analysis [104]. When combined, the quantitative method can verify or not the trends derived from qualitative research [103].

The approach of the questionnaire was based on collecting data from the first-hand user experience. This is called “primary data”. According to Ajayi, “primary data is an original and unique data, which is directly collected by the researcher from a source such as observations, surveys, questionnaires, case studies and interviews according to his requirements” [105]. It is more valid, authentic, objective, and reliable for the research compared to secondary data [106]. Other benefits of this approach are that primary data is real-time and factual and aims to address the specific research problem at hand [105]. Based on the above benefits, the questionnaire was deemed a suitable method to approach the Greek social media community and directly retrieve data for analysis. At the same time, no publicly accessible secondary data on deepfakes and Greece has been found. Considering this, it was necessary to obtain primary data for analysis on the deepfakes landscape in Greece.

The main limitation of the questionnaire as a research method is that it can score a low response rate, despite reaching out to many users. At the same time, the responses may not be honest, as users may not understand the question they are asked, or they have privacy concerns. Providing feedback and clarifications on ambiguous questions is also not possible. There is no direct contact of the researcher with the target audience.

Finally, each participant may interpret the available options in the Likert scale in a different way.

The fourth research question focused on methods used to detect or prevent malicious deepfakes in Greece. The Greek Government has emphasised the importance of cybersecurity [16]. Despite that, there is a growing level of untrust within Greek society towards the Internet and social network platforms and the credibility of information available online [90] [19]. At the same time, Greece is one of the most vulnerable EU states to fake news and misleading content [107]. Lakasas mentions that this could be linked to the mediocre level of education and critical ability when assessing news and information [107]. In the light of this, the need for more cyber safety online was recognised by Facebook and led to novel ideas for combating deepfakes and misinformation online [108] [109].

This investigation utilised the Ellinika Hoaxes, a credibility coalition formed to stop misinformation online in partnership with Facebook. This thesis presented how they detect the deepfakes and their efforts to raise awareness among members of the Greek online community. Also, the questionnaire used in the thesis was distributed among members of the Ellinika Hoaxes to derive observations on their ability to detect deepfakes.

Another tool to detect deepfakes that was analysed was the FB-MSU method used by Facebook AI [110]. The official website of Facebook and the GitHub account with the code overview of the tool were accessed [111] [112].The study aimed to show the technology behind it and describe the strong points of this method against illicit deepfakes.

The analysis on both tools was assisted by the perceptions and views of users from Greece who assessed their confidence level in the Facebook protection framework for deepfakes. Regular Internet and social media users from Greece also expressed their opinions on the effectiveness of the Ellinika hoaxes activities as a non-governmental organisation that addresses online fake content. Again, those observations originated from the analysis of the user responses in the questionnaire.

# 3. Literature review

The chapter aims to critically examine the published studies and articles in the field of deepfakes from a cybersecurity perspective. It first reviews a range of papers and publications that attempt to explain the circumstances that have led to the rise of deepfakes as a cyber threat. It then provides a summary of the major features of deepfakes and gives an overview of how researchers define and categorise them and what technologies are used in their production. In addition, the section provides an indication of the possible implications of malicious deepfakes as featured in the reviewed publications. Also, the literature on detection tools used to mitigate the impact or prevent the spread of deepfakes is investigated. The aim is to show the most cited means of deepfakes detection and provide an assessment of their effectiveness. Finally, this chapter examines the limited studies that focus on testing the human ability to tell apart real visual or audio-visual content from deepfakes.

## 3.1 The rise of deepfakes

Several researchers introduce their studies with the background and factors behind the rise of the deepfakes and the role of social media and the Internet in their promulgation. The research “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security” highlights the role of social media [28]. The authors note that any content, real or fake can go viral. There are more than 3 billion images and 700,000 hours of video shared online daily [29]. In the year 2017, around 95 million photos and videos were shared on Instagram, more than 300 million photos were uploaded on Facebook and, 4,146,600 YouTube videos were watched each day [30]. Facebook recorded more than 4 billion videos every day [31]. Statistics on Instagram show that 59% of the content are photo posts and 14.9% video [31]. McCloskey and Albright refer to the significance of social media for information diffusion [32]. They note that this results in quicker and easier circulation of fake news and deepfakes.

More than 65% of the 2.4 billion Internet users in 2018 accessed their news from social media platforms [33]. The top source for information was Facebook with 43% of the responders using it as their primary news source according to a survey conducted by

Pew Research Centre [34]. Another interesting result in this study is that in more than half of the participants the news reaches them through social media faster than other news outlets. The Reuters Institute Digital News Report notes that the consumption of printed sources of news has fallen in 2020, which provides evidence for the shift to a fully digital future [35]. Later, statistics show that Facebook remains the leader as a news source in the US, with every third American, getting news from this social platform. The next spots are occupied by YouTube (15%), Twitter (12%), and Instagram (11%) [36]. It is expected that Instagram will soon overtake Twitter for news consumption [37].

It can be concluded that social media platforms and the Internet are the dominant environments for content sharing and the mainstream source of news and information. The rise of deepfakes in the cyber domain is linked to the ease of fake imagery generation. The generation of fake content such as Generative Adversarial Networks (GANs) or deepfakes does not require significant experience [32]. The modern Artificial Intelligence (AI) capabilities enable the straightforward creation of deepfakes and fake content [32].

In the article "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking" the authors highlight that in the past editing content required much time and meticulous work. They add that the development of Generative Deep Neural Network (DNN) has changed the way content is synthesised and edited. Most significantly, *DeepFake*, a software utilizing this approach became publicly available in 2018 and was used to create large amounts of fake content [38]. According to the authors, this development led to a surge of deepfakes present online and social media platforms.

At this point, it is notable to consider that deepfakes technology is not only improving but also spreading and becoming accessible to less skilled and sophisticated actors. In the past, those who had access to deepfakes technology were limited to the entertainment industry and sophisticated government organisations [39]. Wittes and Blum note in their book "The future of violence: robots and germs, hackers and drones—confronting a new age of threat" that technologies tend to become cheaper and available to more individuals and mark a rapid pace of growth and diffusion [40]. This means that deepfakes techniques will increasingly be accessible for wider audiences, allowing any individual to create fake content. Parallel to this, tools such as *The Tensorflow* and *Keras* have been democratised with their technical literature widely accessible, and the computational infrastructure required accessed on a low budget [41]. At the same time, deepfakes software has spread to smartphones and the most advanced deepfakes generation models can be accessed by any individual possessing a computer [41].

Another element that contributes to the rise of deepfakes online is the motivation behind their use and the shift towards more malicious use. There are however multiple beneficial uses of deepfakes. For example, their use in the entertainment, gaming, and fashion industries is well documented [39]. Also, education and art have been capitalizing on the possibilities of using deepfakes technology for positive purposes [42]. The use of deepfakes with illicit context was initially limited to the creation of pornography videos of famous personalities or for personal revenge [39]. The latest trend is for politically motivated individuals, groups, and states using deepfakes [39]. This time, the motives have a malicious nature. Specifically, deepfakes can be used for political gains, to influence public opinion and interfere with voting, spread computational propaganda, or cause "artificial panic" within societies and economies" [39].

In the article "One group that's embraced AI: Criminals", Cerulus notes deepfakes have now been embraced by criminals [43]. According to the author, some typical use cases are the generation of fake profiles for phishing scams and impersonations of chief executive officers for corporate frauds. He also mentions fake identity creation which could alleviate money laundering and online frauds. In this context, the motive behind the use of deepfakes has changed drastically. These actors have capitalised on the offensive and damaging capabilities of the deepfakes and have led to their development and spread.

As Hogan mentions in the research "REPLICATING REALITY Advantages and Limitations of Weaponized Deepfake Technology", deepfakes can convincingly distort reality, without the truth being revealed [44]. Some additional qualities present in deepfakes are the difficulty of defending against them. What is more, deepfakes can spread quickly as they are disseminated on social media and the Internet [44]. This means that they reach and impact wide audiences and can shape events before they get detected and stopped. Additionally, as part of an information warfare strategy, they have no clear escalation threshold and enjoy ambiguity under the existing laws, whether national or international. In this regard, they can be used for disinformation and deception [44]. Other scenarios include manipulation of public opinion and shaping the political conversation [44]. Hogan supports that they can cause disruptions to political systems and instigate a crisis within the political and military leadership of an adversary. Also, it is a cost-effective means of undermining trust in political leaders and governmental institutions and foment domestic unrest [44]. Based on the above features, it is becoming clear that state actors such as the USA military became interested in leveraging those possibilities. This interest encouraged the use of deepfakes technology training and development, with more databases created and tested.

All the above considered, malicious motivation such as revenge, theft, frauds, misinformation, political gains, and defamations are catalysts for the rise in the use of deepfakes. The applicability and possibilities for damage are numerous and this makes deepfakes progressively a chosen way of attack for a range of actors including criminals and state-sponsored organisations.

The rise of viral deepfakes can be viewed in terms of a Venn diagram in the Figure 3. Each element corresponds to a different prerequisite for a successful campaign [45]. These components comprise:

a) Tools
b) Social media and the Internet
c) Motivation

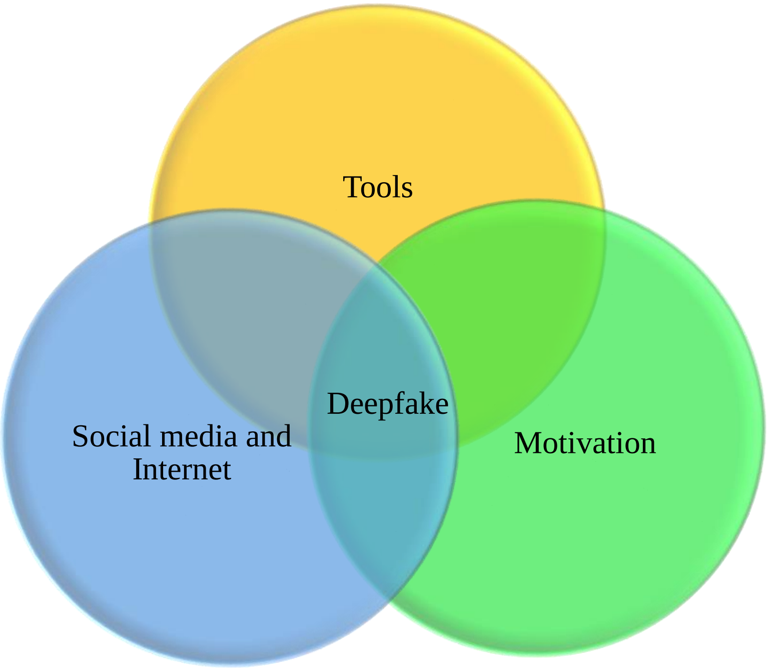


Figure 3. Venn diagram illustrating the constituent elements of a successful viral Deepfake

Each of these elements is necessary for the spread of malicious deepfakes. Tools and deepfakes technology are required to create convincing fake content. Without social media and the Internet as the environment of sharing fake content, deepfakes would not be accessed worldwide and so quickly. Finally, the factor of motivation is highly important, as it is the illicit purposes and applications of deepfakes that made them rise to a global cybersecurity threat.

## 3.2 Deepfakes Definition

Deepfakes have been well defined in the reviewed literature. The paper "Deep Learning for Deepfakes Creation and Detection: A Survey" states that deepfakes are "artificial intelligence-synthesized content" [46]. They can have the form of a face swap, lip-sync, and puppet master. Based on this typology, the authors describe a detailed definition of each type of deepfake. Specifically, face swap deepfakes refer to the superimposition of face images of a target person onto a video of a source person. The final product is a video of the targeted person appearing to do or say things not said or done by the person. Lip-sync deepfake is a video manipulated in such a way so that lip movements are accordant to the audio. Finally, the authors define the puppet master deepfakes as a video in which a person is animated to follow the movements and expressions of the face, eyes, and head of another person. This is referred to as the relation of puppet and master [46].

According to Tolosana, Vera-Rodriguez, et al. A deepfake refers “to a deep learning-based technique able to create fake videos by swapping the face of a person by the face of another person” [47]. In this study, it is obvious that the author highlights the face swap aspect of the deepfakes. Thus, the given definition does not incorporate the cases of image and audio deepfakes. Manipulating the face or body of a person by swapping them with another person and creating a new image is an example of an image deepfake. Voice swapping or changing audio in a recording can also be considered instances of deepfakes, which are not included in the definition.

An alternative definition is given by Hwang. Here, “Deepfakes” include a “broad scope of synthetic images, video, and audio generated through recent breakthroughs in the field of Machine Learning (ML), specifically in deep learning” [48]. This definition is provided by CSET in the report “The Deepfakes: A Grounded Threat Assessment”. The authors of the paper stress that in general the term includes a piece of media modified or generated by ML and non-ML techniques [48]. This is notable because the deepfakes which have been used with malicious intent in disinformation campaigns are not sophisticated [48]. The perpetrators emphasise the scale and massive production and spread of the content, rather than the quality. The study mentions that the most common operational posture exhibited in online influence campaigns is characterised by cheaply crafted audio-visual content and a lack of investment in high-quality deepfakes [48]. This definition given by CSET has two strengths. First, it incorporates a wide selection of possible deepfakes, in terms of the type of the content, but also the technology for manipulation or generation. Second, adding the non-ML methods finds applicability in the real world and the conducted malicious actions with the use of deepfakes. The definition contrasts to mendacity and vagueness.

Another piece of literature in the field of deepfakes is the study by J. Krietzman, Lee, McCarthy, and T. Krietzman titled “Deepfakes: Trick or Treat?” [49]. According to the study, “deepfakes leverage powerful techniques from machine learning and artificial intelligence to manipulate or generate visual and audio content with a high potential to deceive” [49]. Besides this, the authors of the paper provide a detailed background of the origin of deepfakes. They describe deepfakes as the combination of "deep learning and fakes", which was introduced in 2017 by a Reddit user. Many studies follow a similar line when defining deepfakes as a combination of "deep learning" and "fakes. For instance, the "Tech Policy Factsheet: Deepfakes" of the Belfer Centre features the following definition. “The term “deepfake” is a hybrid of the terms "deep learning" and "fake" [50]. It is content produced or manipulated by ML [50].

Articles on news platforms also tend to use a similar definition of deepfakes. Specifically, Metz in an article published for CNN [51], outlines that deepfakes is the combination of the terms "deep learning" and "fake". She delineates that they are false video and audio files that show a real person doing or making statements they did not [51]. Also, Shao in the article “What ‘deepfakes’ are and how they may be dangerous” is again referring to deepfakes in terms of “deep learning” and “fake” [52].

In another article titled “Deepfakes: What they are and why they’re threatening” Johansen notes in a similar pattern that deepfakes blend two words, “deep” and “fakes” [53]. What is more, the author explains that machine-learning algorithms are used to manipulate images and voices to replace the real person’s characteristics with artificial likeness [53]. Similarly, Sample notes the deepfakes are not solely videos, but also images generated from scratch or manipulated audio in the form of “voice skins” or “voice clones” [54].

Additionally, according to Davis, deepfakes can be defined as “synthetic auditory or visual media developed using deep learning, a subfield of machine learning, that appear to be authentic and are often created with the intent of deceiving audiences” [55]. This definition has an important addition compared to the above definitions. Specifically, it incorporates the purpose of the deepfakes use since it highlights the illicit motive behind the use of deepfakes. The aim is to trick societies and spread disinformation and misinformation among individuals [55].

It is worth mentioning that while the definition provided by Davis characterises machine learning as the means of generation or manipulation of visual content, the study also mentions that commonly accessed recording or editing software can be used on images, videos, and audio to produce deepfakes. Thus, deepfakes do not solely rely on ML.

Overall, the term deepfakes has been well defined by relevant studies and most of them find a common ground on the definition of deepfakes. In this thesis, the definition provided by CSET will be used as it embodies a wide range of deepfakes, whether it is an image or, video either with, or without audio. It draws a complete picture of all different deepfakes variations. Another strength of this definition is that it accents both the ML and non-ML nature of deepfakes. Consequently, it does not limit the deepfakes spectrum to only highly sophisticated and technically complex content.

## 3.3 Deepfakes Typology

This section gives a detailed overview of the types of deepfakes, showing the different variations and possibilities of content manipulation and generation.

In the research “Deep Fakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence”, Paris and Donovan provide a deepfakes typology based on the technology used [56]. The spectrum of audio-visual manipulation accommodates deepfakes which are created with “cutting edge, AI-reliant techniques” [56]. Additionally, it includes “cheap fakes” produced with publicly available and cheap software [56]. “Conventional techniques like speeding, slowing, cutting, re-staging, or re-contextualizing footage” are additional methods [56].

Deepfakes that rely on experimental machine learning represent one end of a spectrum of Audio-Visual (AV) manipulation. The deepfake process is both the most computationally reliant and the least publicly accessible means of creating deceptive media [56]. “Cheap fakes” are the other end of the spectrum. The authors claim that in practice both can influence politics and manipulate the evidence [56].

Starting from the most technically sophisticated deepfakes, the spectrum of deepfakes presented in the study contains the following:

a) FakeApp/After Effects: software for manipulating and creating audio-visual fakes.

b) Video Dialogue Replacement (VDR) model: manipulates the source video to appear speaking in different languages [57].
c) Generative Adversarial Networks (GANs).
d) Recurrent Neural Network (RNN); Hidden Markov Models (HMM) and, Long and Short Term Memory (LSTM) Models.

Based on the above typology, the research expounds on the different technologies used for each of the deepfake types. Specifically, methods used for more sophisticated deepfakes include lip-synching, face swapping, voice synthesis and virtual performances. This point has been presented by the several definitions described above. Those methods rely on more technical resources and expertise and skills to generate the final product. Moving on to the deepfakes with decreased technical advancement and complexity, those are the following:

a) After effects/Abode Premiere Pro: software for the creation of motion graphics and visual graphics [58],
b) Sony Vegas Pro: software for audio and video editing [59],
c) Free real-time filter applications: such as Instagram, Snapseed,
d) In-camera effects: effects achieved by the manipulation of the camera or its parts such as time and speed effects, filters, infrared or negative image,
e) Re-labelling/ Reuse of extant video [56].

The study suggests that cheap, publicly available, and easy to download and run software is suitable for the above cheap fakes and less sophisticated deepfakes.

Examples of methods of creation include:

a) Face swapping: Rotoscope: methods of animation from source frames,
b) Speeding and slowing,
c) Face altering/ swapping: with the use of publicly available apps such as SnapChat [56],
d) Lookalikes,
e) Recontextualizing [56].

Those methods are more accessible forms of AV manipulation and not technical but contextual.

The research conducted by Belfer Centre also acknowledges a wide spectrum of deepfakes [50]. The classification depicts the different types of deepfakes, alongside the technique of manipulation used. Based on this, the deepfakes are also classified from “cheap fakes” to more technically sophisticated deepfakes. The “cheap” or “shallow” fakes are created with cheap and commonly available software and lack technical complexity [50]. In this category belong the following:

a) Re-contextualizing: putting existing content into a new manipulated context.
b) Lookalikes: impersonation of others by lookalike hired actors.
c) Speeding and slowing: manipulating the speed characteristics.
d) Face swapping-Rotoscope: replacement of faces between two individuals.
e) Lip-syncing: manipulating the voice and words of a targeted individual.
f) Face replacement: referred also to as deepfake puppetry.
g) Synthetic speech production: generation of voice without a source voice.
h) Face re-enactment and audio modulation: manipulation of a person’s facial and voice characteristics based on pre-defined features.
i) Face generation: production of a new image without a source and face replacement [50].

The less manipulation and amendments to an existing audio-visual or visual content and more content generation with the use of ML, the more sophisticated is the deepfake [50]. The authors of the paper depict the technology used to create the deepfakes. At the same time, the stress is put upon the face swap aspect of deepfakes.

## 3.4 Deepfakes technology

The study by Westerlund titled "The emergence of deepfake technology: A Review" is an analysis of the major features of deepfakes. The author makes use of 84 online articles published during 2018-2019 to present the creation process of the deepfakes and the actors behind them.

A deepfake is created with the use of Generative Adversarial Networks (GANs) [60]. Two artificial neural networks, the generator, and discriminator are combined to deliver AI-manipulated content. The author gives a simple and clear overview of this combination. Both are trained on the same set of images, videos, and sounds, but their role is different. The generator creates new content, by processing thousands of photos [60]. The discriminator determines if a photo is real. The deepfake created by the generator is a photo similar to the original, but not an exact copy of it. If the forgery is realistic enough, the discriminator will not detect that it is fake [60]. The portrayal of the deepfakes technology by Westerlund is simplified and depicts the basic idea of the deepfakes creation technology. It does not go deep into details and is not supported by a visual example, which would make the topic more approachable.

As stated by Deeptrace, the phenomenon of deepfakes is evolving rapidly [14]. This process is intensified by the growing accessibility and development of tools that enable even non-experts to create deepfakes [14]. A notable development in this field is the deepfake-generating system called FaceSwaping Generative Adversarial Network (FSGAN) [61]. It was created by Israel’s Bar-Ilan University. The major difference in this technology compared to the previously reviewed is that FSGAN does not require the step of software training to produce a deepfake [62]. In the light of such developments, the deepfakes creation will continue to rely on less technical know-how [62].

The authors conclude that their work makes face swapping and re-enactment accessible to non-experts. While the above papers depict the sophisticated technology and steps of the deepfakes creation process, it is worth mentioning that deepfakes can also be created with publicly available software. Many online guides show step by step how to create a simple deepfake image or video, with the use of open software. These include the “*walkthrough*” from Techwiser, which gives a detailed explanation of the DeepFaceLab to create deepfakes [63]. It starts with instructions on how to download the software and then it proceeds on each step necessary to produce a deepfake. The main steps are the following:

1) Adding video files to the project. Two files are required, namely one source and one destination.
2) Extraction of frames from the videos: this step allows to extract frames from both source and destination video.
3) Extraction of faces from frames: once the frames are collected, the next step is to extract faces from those frames, both source, and destination video.

4) Training: the software needs to be trained to recognise and link facial expressions. This way, the original and generated face will resemble each other. The result is to swap those two convincingly.
5) Faces conversion into frames: the faces are pasted into the final video frames.
6) Frames conversion into a deepfake video: the assembled frames result in a deepfake video with a face swap [63].

In another article, DFBlue explains in even simpler and fewer steps how to use the DeepFaceLab to create deepfakes [64]. Ravindu Senaratne, in the article “Make Your Own Deepfake Video in a Few Easy Steps”, provides another way to create deepfake videos. [65] Specifically, he relies on First Order Motion Model for Image Animation by Aliaksandr Siarohin [66]. The process seems not as complex as the previously mentioned tool and enables the creation of deepfakes quickly and cheaply. The major elements required are the suitable size of images used in the deepfake content and the successful execution of commands, provided on the website [65].

The Deepfakes Web is one of the available applications for deepfakes creation [67]. It enables the creation of deepfake videos by subscription. Also, FakeApp is another software that enables face-swapping [68].

## 3.5 Deepfakes impact

The most valuable input in terms of the impact of deepfake on cybersecurity is presented in the study “The State of Deepfakes: Landscape, Threats, and Impact” by Deeptrace” [14]. This offers an in-depth analysis of the effects of malicious deepfakes on individuals, politics, and businesses.

Specifically, it delineates that deepfakes have affected many women and famous personalities in the context of creating pornographic content involving them without their consent. The phenomenon is spread on a global scale and intensified by popularity and viewership. Moreover, online software such as DeepNude will only increase and simplify the process of deepfake pornography spreading online [14].

Politicians and governments are also increasingly damaged by deepfakes. In more detail, the researchers explain that deepfakes can be used in campaigns to undermine democratic processes and manipulate elections. They can distort the political discourse, delegitimise political figures, and represent political news in a false and manipulated manner [14]. The authors support their statements with some real-world examples of deepfakes destabilizing the political sphere. The most striking example is the case of Gabon in 2018 in which the government aimed to end speculations surrounding the health and absence of President Bongo. The video delivered to the public was believed to be a deepfake. This sparked unrest and lead to a coup d'état against the Government. In summary, deepfakes undermine the objectivity of videos featuring political figures, create manipulations of reality, and disrupt everyday politics.

The study by Deeptrace has a separate section discussing the impact of deepfakes on the cybersecurity landscape. They introduce new cyber threats and expose new points of attack and unauthorised access. The study foresees two major cyber threats. Firstly, cyber frauds, espionage, and infiltration based on fake digital identities. Secondly, the threat of impersonations and frauds against individuals and businesses with the use of synthetic audio-visual content is stated. The first case refers to the use of synthetically generated images of people to create fake social accounts such as Maisy Kinsley and Katie Jones [13]. Both were fake personas with profiles on LinkedIn and Twitter respectively and were involved in espionage campaigns [13]. [14]. In parallel, synthetic voice cloning can be leveraged as a new means of online frauds and impersonations of Chief Executive Officers (CEO) in the corporate world. Altogether, deepfakes add to the tools available to cybercriminals with new means for social engineering and frauds [69].

In another study conducted by Westerlund, deepfakes are portrayed to have a beneficial role for many industries. Entertainment and media industries, e-commerce, fashion, health, games industry, education, and material science are mentioned [60]. At the same time, the study explored the dark side of the deepfakes as a threat to modern societies, businesses, and political systems. The analysis on both sides of the deepfakes technology is valuable and parallel to the notion that technology has both a bright and dark side and opportunities for individuals, organisations, and governments [49].

Westerlund states that:

a) "deepfakes put pressure on journalists struggling to filter real from fake news.

b) They threaten national security by disseminating propaganda and interfering in elections.
c) They hamper citizen trust toward information by authorities
d) The raise cybersecurity issues for people and organisations.” [60]

One of the key consequences of deepfakes is the phenomenon of "information apocalypse" or "reality apathy". Faced with constant exposure to fake content and misinformation, individuals tend to perceive any information as not true and deceiving [60].

Additionally, the research emphasises the cybersecurity threats posed by the deepfakes. In detail, the author mentions that deepfakes can be used in corporate frauds, cybercrimes, and market and stock manipulation. The author mentions that deepfakes can be used for the impersonation of high executives to trick into financial frauds or data leaks. In addition to this, the study raises the issue of malicious scripts, websites mining cryptocurrencies from their visitors, and crypto-jacking threats for deepfakes hobbyists.

This is a useful addition to the debate on deepfakes, as the cybersecurity aspect is often not covered by researchers. To support this view, the study by J. Krietzman, Lee, McCarthy, and T. Krietzman titled “Deepfakes: Trick or Treat?” provides a valuable contribution [49]. This piece of literature takes under its scope the impact of deepfakes on individuals, organisations, and governments. It highlights the view that like any technology, deepfakes are also used with good and bad intentions. From the point of view of an individual, deepfakes have an enjoyable and entertaining nature, enabling people to swap their faces and produce deepfakes for fun. However, with malicious intent manipulating content with the use of AI incorporates threats to privacy and identity or leads to defamation, fraud, and reputational damage [60].

The authors stress the financial and reputational damage to organisations caused by malicious deepfakes used in frauds and trickery. This may involve the manipulations of markets and result in unexpected fluctuations to share prices. Among the benefits of deepfakes, they state that they are used in the entertainment and fashion industries, frequently used for voice dubbing actors in film production. Finally, they focus on the role of governments. They mention the high value of deepfakes technology in communication and broadcasting messages with an awareness value for the public.

However, malicious deepfakes can mislead and disinform societies. They can be used as means to manipulate public opinion and thus impact important decisions and political events, such as elections. Propaganda and confusion of the public can eventually undermine the trust in public institutions and heads of government.

To summarise this section, the study shows the deceptive and destructive side of the deepfakes on individuals, organisations, and governments. Despite that, it does not refer to the impact on the cybersecurity of the three groups.

## 3.6 Deepfakes detection technologies

The previous section illustrated the serious implications that deepfakes have on privacy, the cybersecurity of users and businesses, and the integrity of information. As a result of this threat, a range of detection methods have been developed, which are examined in this section.

### 3.6.1 Recurrent Neural Networks (RNN)

The paper "Deepfake Video Detection Using Recurrent Neural Networks" describes a recurrent deepfake video detection system [41]. Guera and Depl explain that the system is based upon the following components:

a) Convolutional Neural Network (CNN) for features extraction for each frame.
b) Long Short-Term Memory (LSTM) for sequence processing.

The overview of the detection methods can be seen in the following Figure 4.

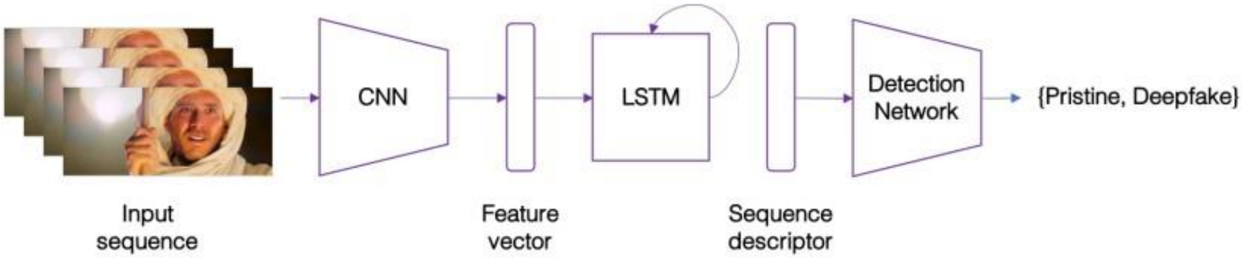


Figure 4. Recurrent Network for Deepfake Detection

CNN applies filters to an input video and creates a map of the features present [70]. A sequence of frames with those features is constructed and passed through the Convolutional Long Short Term Memory (Conv-LSTM) It processes the input and produces a sequence descriptor that computes the probabilities of the frame sequence being either pristine or deepfake [41]. The method was evaluated on 600 videos, 300 real videos randomly picked from the Hollywood Human Actions (HOHA) dataset, and 300 deepfake videos collected from various video-hosting webpages [71]. The results on the accuracy of this technique can be seen in the following Table 1.

Table 1. Deepfakes detection accuracy of RRN method

| **Model** | **Training accuracy** | **Validation accuracy** | **Test accuracy** |
|---|---|---|---|
| Conv-LSTM<br>20 FRAMES | 99.5 | 96.9 | 96.7 |
| Conv-LSTM<br>40 FRAMES | 99.3 | 97.1 | 97.1 |
| Conv-LSTM<br>80 FRAMES | 99.7 | 97.2 | 97.1 |

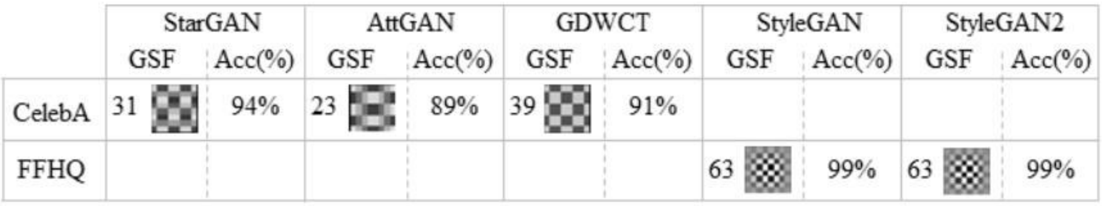
The researchers used subsequence of 20, 40, and 80 frames without frame skips. The accuracy of the method is around 97%, for the tested video lengths.  According to Guera and Depl, this means that their method can detect with high accuracy if a video was manipulated or not with just 2 seconds of video duration [41]. In their methodology, 40 frames for videos are sampled at 24 frames per second, thus the video lasts no longer than 2 seconds.

### 3.6.2 Generative Adversarial Networks (GAN) Discrete Cosine Transform (DCT) Anomalies

In the paper “Fighting Deepfakes by Detecting GAN DCT Anomalies”, the authors Giudice, Guarnera, and Battiato propose a deepfake detection method that analyses

Generative Adversarial Network (GAN) Specific Frequencies (GSF) [72]. They explain that the signature of the generative process is embedded into those frequencies and can be calculated with the analysis of the Discrete Cosine Transform (DCT) coefficients [72]. The main benefits of this approach are the limited computational power and time spent on training the system and the identification of GAN-specific deepfake architecture based on anomalies patterns detected [72].

Figure 5 depicts some basic features of the evaluation process for the GAN DCT approach. Specifically, the datasets CelebA and Flickr-Faces-HQ Dataset (FFHG) are used for authentic images. StarGAN, AttGAN, GDWCT, StyleGAN, and StyleGAN2 are GANs architectures used to generate deepfake images. More details provided by the authors of the research mention that 700 images were used for testing purposes and 200 images for training. The results demonstrate a general high impressive percentage of accurate detection of deepfakes.



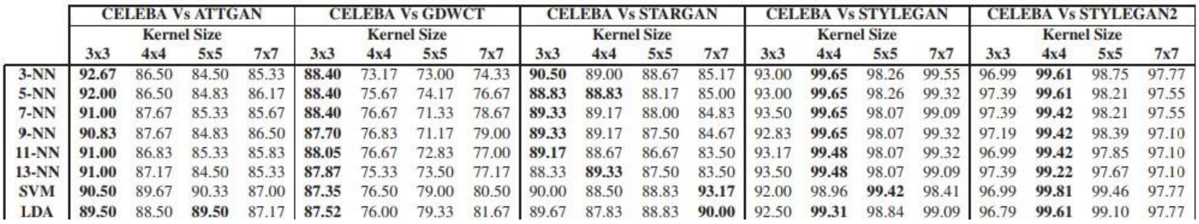
| | StarGAN | | AttGAN | | GDWCT | | StyleGAN | | StyleGAN2 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | GSF | Acc(%) | GSF | Acc(%) | GSF | Acc(%) | GSF | Acc(%) | GSF | Acc(%) |
| CelebA | 31 | 94% | 23 | 89% | 39 | 91% | | | | |
| FFHQ | | | | | | | 63 | 99% | 63 | 99% |

Figure 5. Accuracy results for detection with the GSF [72]

### 3.6.3 Convolutional Neural Network Methodologies

The detection method described in the chapter relies on the Expectation Maximisation (EM) algorithm. The algorithm is used to mathematically capture the correlations

between pixels in the image [73]. The researchers behind this method tested and trained six different datasets of images. CELEBA was the source for authentic content [73]. Deepfakes were generated with the use of five GANs, namely STARGAN, STYLEGAN, STYLEGAN2, GDWCT, ATTGAN [73]. The accuracy results are reproduced in Figure 6 below.



|  | CELEBA Vs ATTGAN |  |  |  | CELEBA Vs GDWCT |  |  |  | CELEBA Vs STARGAN |  |  |  | CELEBA Vs STYLEGAN |  |  |  | CELEBA Vs STYLEGAN2 |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | Kernel Size |  |  |  | Kernel Size |  |  |  | Kernel Size |  |  |  | Kernel Size |  |  |  | Kernel Size |  |  |  |
|  | 3x3 | 4x4 | 5x5 | 7x7 | 3x3 | 4x4 | 5x5 | 7x7 | 3x3 | 4x4 | 5x5 | 7x7 | 3x3 | 4x4 | 5x5 | 7x7 | 3x3 | 4x4 | 5x5 | 7x7 |
| **3-NN** | **92.67** | 86.50 | 84.50 | 85.33 | **88.40** | 73.17 | 73.00 | 74.33 | **90.50** | 89.00 | 88.67 | 85.17 | 93.00 | **99.65** | 98.26 | 99.55 | 96.99 | **99.61** | 98.75 | 97.77 |
| **5-NN** | **92.00** | 86.50 | 84.83 | 86.17 | **88.40** | 75.67 | 74.17 | 76.67 | **88.83** | **88.83** | 88.17 | 85.00 | 93.00 | **99.65** | 98.26 | 99.32 | 97.39 | **99.61** | 98.21 | 97.55 |
| **7-NN** | **91.00** | 87.67 | 85.33 | 85.67 | **88.40** | 76.67 | 71.33 | 78.67 | **89.33** | 89.17 | 88.00 | 84.83 | 93.50 | **99.65** | 98.07 | 99.09 | 97.39 | **99.42** | 98.21 | 97.55 |
| **9-NN** | **90.83** | 87.67 | 84.83 | 86.50 | **87.70** | 76.83 | 71.17 | 79.00 | **89.33** | 89.17 | 87.50 | 84.67 | 92.83 | **99.65** | 98.07 | 99.32 | 97.19 | **99.42** | 98.39 | 97.10 |
| **11-NN** | **91.00** | 86.83 | 85.33 | 85.83 | **88.05** | 76.67 | 72.83 | 77.00 | **89.17** | 88.67 | 86.67 | 83.50 | 93.17 | **99.48** | 98.07 | 99.32 | 96.99 | **99.42** | 97.85 | 97.10 |
| **13-NN** | **91.00** | 87.17 | 84.50 | 85.33 | **87.87** | 75.33 | 73.50 | 77.17 | 88.33 | **89.33** | 87.50 | 83.50 | 93.50 | **99.48** | 98.07 | 99.09 | 97.39 | **99.22** | 97.67 | 97.10 |
| **SVM** | **90.50** | 89.67 | 90.33 | 87.00 | **87.35** | 76.50 | 79.00 | 80.50 | 90.00 | 88.50 | 88.83 | **93.17** | 92.00 | 98.96 | **99.42** | 98.41 | 96.99 | **99.81** | 99.46 | 97.77 |
| **LDA** | **89.50** | 88.50 | **89.50** | 87.17 | **87.52** | 76.00 | 79.33 | 81.67 | 89.67 | 87.83 | 88.83 | **90.00** | 92.50 | **99.31** | 98.84 | 99.09 | 96.79 | **99.61** | 99.10 | 97.77 |

Figure 6. Detection accuracy results for EM algorithm [73]

The accuracy is provided in a comparison format between CELEBA and each of the GANs. Different kernel sizes (filters applied by CNN on the input image to extract features) are also tested and different classifiers are used. The K-Nearest Neighbour (KNN) of kernel size set at 3, 5, 7, 9, 11, and 13, Linear SVM and Linear Discriminant Analysis (LDA) are used [73].

From the above results, the following conclusions can be reached:

a) In comparison between CELEBA and ATTGAN, the highest accuracy of almost 93% is detected with a kernel size of 3x3 and 3-NN.
b) CELEBA compared to GDWCT showed the maximum detection percentage with a kernel size of 3x3 and 3NN, 5-NN, and 7-NN. The percentage was slightly higher than 88%.
c) In comparison to CELEBA and STARGAN, the detection rate reached 93.17%. It was obtained with linear SVM and kernel size of 7x7.
d) The detection rate for CELEBA compared to STYLEGAN was 99.65% with 3-NN, 5-NN, 7-NN, and 9-NN and a kernel size of 4x4.
e) Finally, CELEBA Vs STYLEGAN2 showed the highest accuracy with linear SVG and 4x4 as the size of the kernel. It skyrocketed to 99.81% of correct detection.

A closer look at the above table makes it clear that in general the rates of accurate detection of deepfake images based on the proposed method are high. As a future step, the method could also be tested and trained on videos to tell apart deepfake from authentic content.

### 3.6.4 Biological signals

The field of deepfakes detection tools is being constantly enriched with more novel methods. Binghamton University and Intel Corporation created a detection method based on the disentanglement of generation residuals with biological signals [74].

Figure 6 below illustrates the process of detecting a deepfake based on biological signals. Specifically, the source video which is genuine content (a) is transformed into a deepfake (c) with the use of generators (b). Each of the models, corresponding to point c, have their residuals after their generation [74].

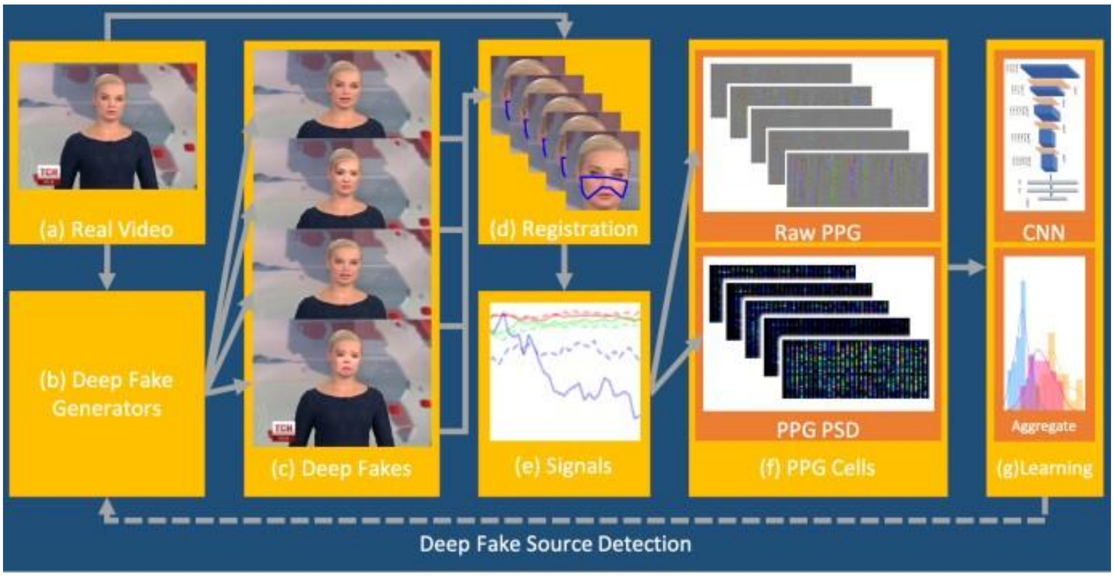


Figure 7. Deepfake Source Detection with Biological Signals [74]

The authors explain that a detector is used to extract faces. Then, the Regions of Interest (ROIs) of each face that have the most stable Photoplethysmography (PPG) signals, such as heartbeat and blood flow are chosen. The area of focus in the region between mouth and eyes, assigned in Figure 6, point d.  All signals are linked and give the final ROIs of the faces. These PPG cells combine several raw PPG signals and their power spectra, extracted from a fixed window. This can be seen in point f of Figure 6 [75]. The final step is a combination of PPG cell training and window prediction aggregation. The authors conducted experiments with the use of different window lengths and different datasets of deepfakes. They conclude that the novel method they tested achieved 93.39% accuracy on the dataset of FaceForensics++ videos [74]. As the authors note, the main strength of their approach is that it has a high rate of accuracy, and it is used not only to detect the deepfakes from real content but also to predict the source generative model of a deepfake.

### 3.6.5 Phoneme-Viseme Mismatches

A joined research of Agarwal, Farid, Fried, and Agrawala describes a deepfake videos detection method from Phoneme-Viseme Mismatches [76]. Their approach foresees that deepfake videos have occasional mismatches in the viseme-phonemes.

The authors explain that phonemes are the units of sound, while visemes refer to the dynamic of the mouth shape and constitute the visual part when a phoneme is articulated. This is illustrated in Figure 7. They bring about the following explanations. The viseme M, B, P, corresponds to the phoneme of the words like “mother”, “brother” and “parent” [76]. The pronunciation of "chair" (CH), "jar" (JH), or "shelf" (SH) can be seen on the last image from the right. Based on the above, there are sets of unique mappings of viseme and phoneme.  For each word, the mouth can be shaped differently.

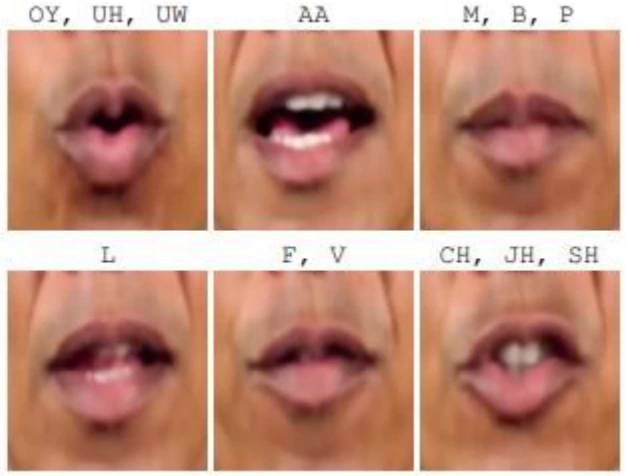


Figure 8. Viseme-Phoneme mappings [76]

For the research, a dataset of Audio-to-Video (A2V), Text-to-Video with short statements (T2V-S), and long statements (T2V-L) were used. In the case of A2V, the deepfake is a video of a real person with a manipulated voice. Also, the T2V, whether short or long, is a deepfake showing a real person with a mouth synchronised to say a manipulated set of words.

The process of measuring visemes is done in three different ways. The researchers extract the viseme manually from six video frames around the occurrence of the phoneme group of M, B, and P.  A second approach is to extract the visemes from the same frames of phonemes automatically. A third approach is the measurements with the use of a convolutional neural network (CNN). Each of the three methods is used to classify if the mouth in a deepfake video is open or closed and if the pairing of viseme-phoneme is correct [76].  It must be noted that the CNN technique is person-specific and has been trained only on videos of Obama. Thus, the performance of this method is expected by the authors to be much better on the videos of Obama contrary to videos showing a different person [76]. The following Figure 9 shows the results of the manual detection with the use of a graph.

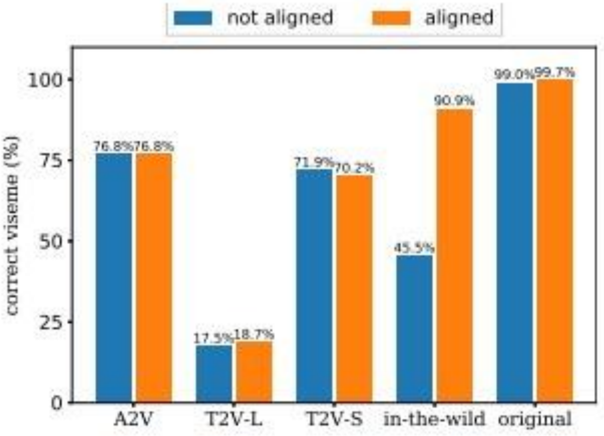


Figure 9. Percentage of correct Viseme-Phoneme pairings (manual deepfakes detection) [76]

For each dataset, the blue column represents the percent of MBP phoneme occurrences where the correct viseme is detected, before the audio to video alignment. The orange column shows that percentage after the audio is aligned to the video. The results show a high percentage of correct detection of deepfake videos based on the accurate viseme and phoneme mapping. Apart from the deepfakes in the wild, the alignment of audio to the video did not significantly affect the percentage. The detection success percentages for the automatic and CNN-based detection methods are shown in Table 2.

Table 2. Percentage of correct Viseme-Phoneme pairings (automatic and CNN deepfakes detection) [76]

| **DATASET** | **PROFILE** | **CNN** |
|---|---|---|
| Original | 99.4% | 99.6% |
| A2V | 96.6% | 96.9% |
| T2V-L | 83.7% | 71.1% |
| T2V-S | 89.5% | 80.7% |
| In-the-wild | 93.9% | 97.0% |

The researchers note that the accuracies are computed at a fixed threshold of 0,5% as the average false alarm rate of interpreting a closed mouth as open and vice versa. Both methods have approximate accuracy percentages in the original and A2V datasets of deepfakes. The most significant deviations are observed in the T2V-L and T2V-S datasets. Deepfakes in the wild are also detected with different levels of accuracy by the two methods. Specifically, the CNN method has a higher success rate in the category of deepfakes in the wild, but it is less accurate compared to the profile detection in both T2V deepfakes.

The study presents a comparison of all three methods in detection of the deepfakes, as the length of the deepfake video progresses from 10 to 30 seconds (Figure 10). Each curve represents a different dataset of deepfakes tested by the researchers (original, A2V, T2V-L, TV2-S, in-the-wild). The study also reminds us that the detection of the deep fakes relies on finding mismatched MBP phoneme to viseme [76]. The detection accuracy increases as the duration of the video maximise to 30 seconds. This applies to all three approaches and for all involved datasets apart from `the original one. Another point to note is that manual detection hits higher percentages of accuracy compared to profile detection. The study provides the following results when the video length reaches 30 seconds. The manual detection for original, A2V, and T2V-S datasets sets foot on 96.0%, 97.8%, and 97.4%, while the automatic profile technique records a slight drop to 93.4%, 97%, and 92,8% respectively. The CNN method registers 93,4% and 97,8% for the original and A2V datasets, while for the T2V-S dataset, the percentage of correct mismatches detection falls by circa 7% and hits 81% [76].

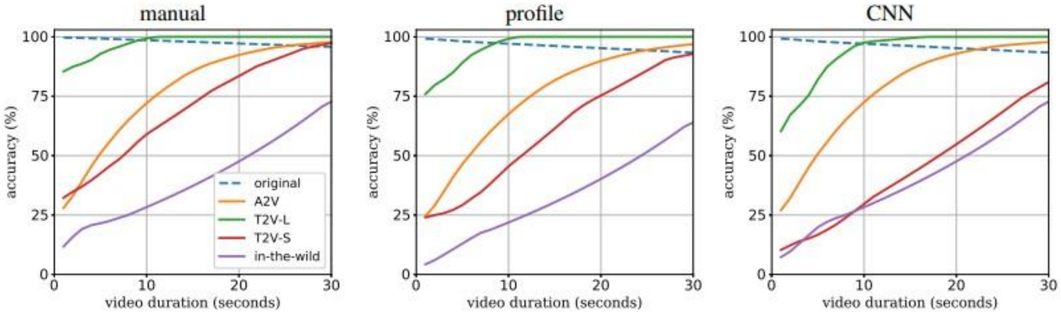


Figure 10. Comparison of the accuracy percentages in correlation to a deepfake video duration [76]

Finally, the automatic detection techniques are tested for their robustness after laundering the deepfake videos. The researchers conducted recompression and resizing on the videos. The profile technique has an average accuracy of 90.46% after recompression, while 83.8% after resizing. The CNN technique detects viseme-phoneme mappings with an average accuracy of 99.32% after recompression and 89.92% after resizing the deepfake video [76]. The results compared to the ones that can be seen in Figure 8 reveal that the two video laundering operations have a noticeable impact on the profile detection technique, while the CNN method is not notably affected.

### 3.6.6 Facial movements

The study “Protecting World Leaders Against Deep Fakes” proposes a forensic method for detecting deepfakes based on facial behaviour analysis [77]. It uses the toolkit OpenFace2 to extract facial and head movements in a video. The authors explain that those movements are unique to everyone. Thus, when comparing the extracted expressions from deepfakes, irregularities and disruptions of those patterns become visible [77]. The impersonators' expressions are different compared to the source individual [77]. Also, in the case of just lip-syncing and not face-swap, the mouth looks disconnected from the rest of the face [77].

The focus is put upon national leaders and high office candidates. This choice is supported by the threat that the deepfakes pose to democracies and elections [77].The method was tested with the use of content showing Hillary Clinton, Barack Obama, Bernie Sanders, Donald Trump, and Elizabeth Warren [77]. The method is tested to tell apart the 10-second clip and full video segment of a real politician from the fake contents of a random person, comedian or impersonator, face swap, lip-sync, and puppet master [77]. Also, one test compares 190 features of a face, while the second 29 features [77]. Based on the accuracy results provided in the study, it becomes visible that the more features incorporated in the comparison the lower the accuracy of the detection. The lowest accuracy is detected in the lip-syncing content, which manipulates only the mouth region. Also, the length of the source video matters. The more seconds it features, the higher the accuracy of the tool in detecting correctly the deepfake.

The tool was also tested for its robustness to recompression. It is not pixel-based. As a result, laundering cannot impact the accuracy of the detection [77]. Finally, the method is not robust to different contexts, in which the same person can have different facial expressions and talking style [77]. The authors tested the accuracy of their analysis by using weekly addresses of Barack Obama and interview videos. The model could not capture all the features, which cause a significant drop in accuracy [77]. In this light, a bigger and more diverse dataset is a challenge for this model.

The paper “Detecting Deepfake Videos: An Analysis of Three Techniques” compares the effectiveness of deepfakes detection by convolutional LSTM, eye blink detection, and grayscale histograms pursued [78]. The Long-term Recurrent Convolutional Networks (LRCN) has been described above in the research “Deepfake Video Detection Using Recurrent Neural Networks” [41].

One of the main challenges for deepfakes creation is the reproduction of the eye blinking [78]. According to the authors of the paper, the lack of this feature may indicate a deepfake. As seen in the Table 3, the fake videos have a reduced number of eye blinks compared to the real content.

Table 3. The average number of eye blinks in 10 seconds in real life, real video and fake videos [53]

| **Real life** | **Real videos** | **Fake videos** |
|---|---|---|
| 3.4 | 4.8 | 2.2 |

The above statement is present in the research **“**In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking” [38]. The general picture of the method is described in the following Figure 11. The process starts with extracting faces from the input frames, as shown in point a. If the head movements and changes in face placement are present, the authors used 2D face alignment. The face must be in the centre of the image and the eyes need to lie horizontally [38]. After the face alignment, the areas surrounding the eyes are extracted into a new sequence of frames. This is shown in point b of Figure 11. This region is passed into the LRCN. The methodology of LRCN is illustrated in point c. The main components are feature extraction, sequence learning, and state prediction [38]. For feature extraction, CNN is issued, while RNN combined with LSTM is responsible for the sequence learning [38]. The final stage is the prediction of eyes open or closed. It is denoted as 0 and 1 respectively [38]. For the final state prediction stage, the output of each RNN neuron is further sent to the neural network consists of a fully connected layer, which takes the output of LSTM and generates the probability of eye open and close state, denoted by 0 and 1 respectively.

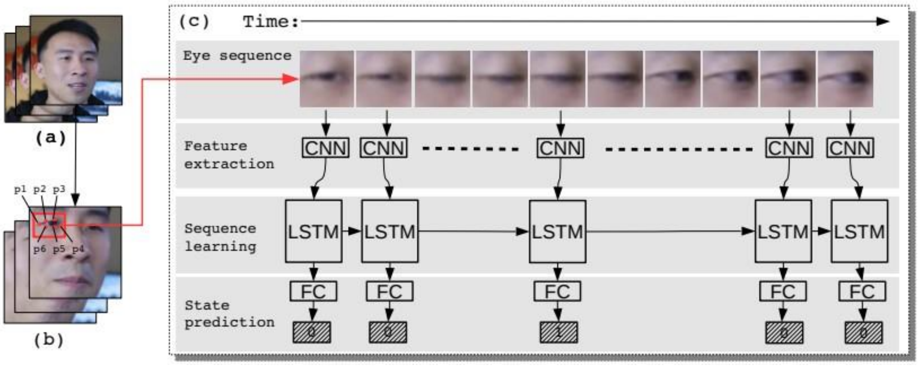


Figure 11.Deepfakes detection with the LRCN method

The researchers formed their dataset for experiments on this method, the Eye Blinking Video (EBV) dataset. They used 50 videos with 30 seconds duration and at least one blinking detected [38]. The results of the tests applied to the dataset are shown in Figure 12.

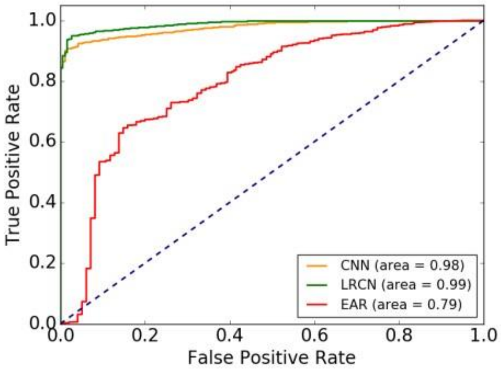


Figure 12. ROC curve comparisons of CCN, LRCN, and EAR for eye state detection [38]

Based on the Receiver Operating Characteristic (ROC) curve used by the researchers; it is visible that the LRCN scored the highest percentage of all models tested. It noted 0.00 compared to 0.98 of CNN and 0.79 of the Eye-Aspect Ratio (EAR) algorithms. The researchers make an important observation on these results. Specifically, the LRCN can memorise the previous state of the eye, while the CNN does not [38]. Thus, if the previous state contained blinking eyes, then the LRCN will predict that the next few frames will include eyes open. Thus, the LRCN is believed to be smoother and more accurate, while CNN can be prone to confusion [38].

The third method analysed in the paper "Detecting Deepfake Videos: An Analysis of Three Techniques" relies upon the grayscale histograms. The paper makes direct reference to the article of McCloskey and Albright title "Detecting GAN-generated Imagery using Colour Cues" [79]. Their method emphasises the way the GANs generator system treats the colour in GANs in a different way compared to real camera images [32]. The saturation of pixels and colours of the image are the components that were tested. The results achieved by the researchers are the following:

a) The pixel intensities approach provides good discrimination between GANs and camera images.
b) The method is not fully effective when comparing fully GAN-generated imagery from natural images.
c) The colour image forensic approach is less effective compared to the pixel saturation-based method [32].

These conclusions can be viewed in Figures 13 and 14, which describe the performance of the saturation statistics and the colour image forensic methods.

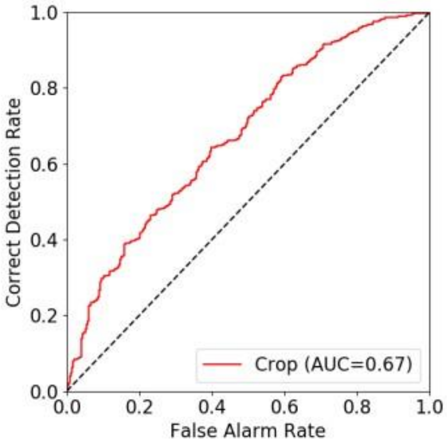


Figure 13. Accuracy with pixel saturation approach [32]

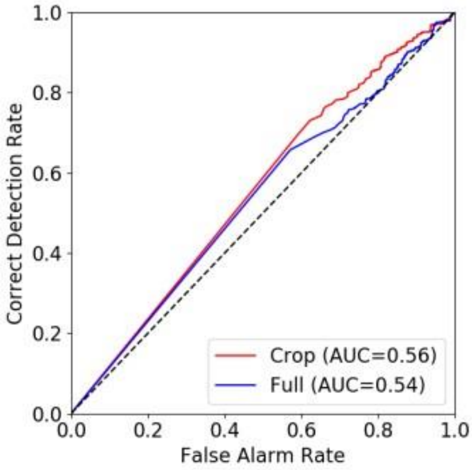


Figure 14. Accuracy with colour image forensics [32]

Pishori, et al. support that a grayscale histogram can be used instead of the colour-based method. This eliminates the computational challenges deriving from colour dimensions [78]. The comparison of the three above approaches produced the following results. Table 4 describes the accuracy, validation loss, and validation accuracy for the selected models.

Table 4. Comparison of the accuracy, validation loss, and validation accuracy [78]

| **Model** | **Accuracy** | **Validation Loss** | **Validation Accuracy** |
|---|---|---|---|
| CNN+RNN | 82.20% | 1.6847 | 82.81% |
| Eye Blink Detection | 81.67% | 0.4762 | 81.67% |
| Grayscale Histogram | 85.71% | 0.5927 | 81.32% |

The researchers incorporated optimisation methods to maximise the accuracy of all compared models. Also, a dataset of 50GB was used for training purposes which according to the authors led to lower accuracy results. From the table, the following conclusion can be derived. The grayscale histogram reached the highest percentage of accuracy, while the eye blink detection method was the least successful in detecting deepfakes.

The authors note that their comparisons show how rapid is the development of new detection methods. Thus, models such as the eye blink detection have lost effectiveness [78]. This may mean that those methods will become quickly obsolescent and no longer effective. They also view that the grayscale histogram method while reached the highest accuracy has a potential for increased effectiveness with proper training and time and a larger dataset. The drawback is that it requires the generation of grayscale histograms for each video, which is a computational challenge [78].

The overviewed research provides a valuable insight into the accuracies of deepfakes detection methods, based on systems and architecture differences of the selected models. It is a direct approach to make the benefits of some models more visible. It also proves clearly that some methods are falling into disuse while others require more research and resources. The approaches for deepfakes detection should not be limited to one single technique, but combinations of them could have beneficial results in detection accuracies.

In short, there is a growing number of methods developed to detect deepfakes which aspire to achieve the highest possible accuracy. Researchers and companies are perceiving deepfakes as a new wave of cybersecurity threats. Their models and techniques take advantage of the special characteristics of authentic videos and images, which make them distinct from AI-manipulated content. This AI arms race is preparing a line of defence against both less and more sophisticated deepfakes, before their online spread becomes noticeable and their impact on individuals, organisations and governments becomes alarming. The analysed techniques show high rates of detection accuracies and give a promise of successful fight against the abuse of deepfakes online.

## 3.7 Human ability to detect deepfakes

Human perception differs from the ability of the algorithms and machine learning models in the detection of deepfakes. A study "Human detection of machine manipulated media" forms an experiment that tests the ability of over 15,000 individuals to detect deepfakes [80]. The researchers note that after being exposed to 10 images the subjects began to learn to detect deepfakes and their accuracy increased. Furthermore, they discovered that individuals who participated in the survey via mobile phone learned faster when compared to those who detected the deepfakes with the use of a computer. Their main assumption is that the zoom feature in mobile phones allows closer investigation of each image.

Another paper that studies the human ability to detect deepfakes is the “Human Perception of Audio Deepfakes” by Muller, Markert, and Bottinger [81]. Their experiments compared the accuracy of 200 individuals and an Artificial Intelligence algorithm in detecting audio deepfakes. The country of origin, age, and IT experience of the participants varied [81]. The results of the experiment highlight that the algorithm trained and used to detect audio deepfakes were superior to humans in terms of ability to detect deepfakes [81]. At the same time, younger participants up to 35 years old scored higher accuracy levels compared to the older participants [81]. Finally, the Information Technology (IT) skills did not impact the ability to detect audio deepfakes [81]. Despite this the researchers note that there is some bias in their study towards young male German individuals with IT skills and experience. What is more, they stress that humans and AI-algorithms abilities are complementary. Specifically, “where one performs strongly, the other fails, and vice versa” [81]. Thus, they suggest that the detection of deepfakes may follow a hybrid pattern that combines both AI algorithms and human ability to detect deepfakes.

A similar conclusion is made by Korshunov and Marcel in their research “Deepfake detection: humans vs. machines” [82]. They used 120 videos from the Facebook deepfake database provided in Kaggle's Deepfake Detection Challenge 2020 [82]. Half of them were authentic and half deepfakes [82]. 60 participants were asked to decide whether the shown content was deepfake or not [82]. Their performance was compared with the results of two detection methods both relying on neural networks [82]. The findings of the experiment show that humans detect easily deepfakes which machines find difficult to spot, while they struggle with deepfakes easily detectable by the detection algorithms [82].

The study “Comparing Human and Machine Deepfake Detection with Affective and Holistic Processing” shows that human detection ability could be a promising defence line against the threat of deepfakes [83]. More than 5,000 participants were challenged to detect deepfakes. The findings of the experiments show that the human and computer vision model have different results based on the types of videos considered [83]. Blurry, very dark, and grainy videos are easier detected by machines, while humans detect easier lip sync videos, as well as attention check and political videos [83]. At the same time, both machines and humans perform equally well when they are challenged with manipulated images on a standard resolution and not videos [83]. Finally, the research delineates that the human abilities to generalise and to do holistic visual processing of the faces are strengths of humans in detecting deepfakes. This, as suggested by the authors, could be capitalised on the existing computer vision models for deepfake detection. They also denote that social media content moderation is probably more precise when conducted by humans compared to machines.

The human ability to detect manipulated content is used by social media companies such as Facebook to detect policy violations on the platform. Facebook has outsourced the content moderation tasks [84]. Greece was chosen as one of the few countries to hold a content moderation hub for Facebook. The main task of the content moderators is to determine whether the content, publicly available on Facebook, violates the Community standards and thus should be removed [85]. Facebook policy bans those videos and images that are edited or synthesised with the use of artificial intelligence or machine learning. The reasoning is that they have a misleading and manipulative effect on users [86]. Thus, judging and removing such content is also part of the moderators’ responsibilities. The social media company is developing an innovative idea for fact-checking online content by assigning this task to the Greek group Ellinika hoaxes, detailed in Annex C [87].

The European Union Agency for Cybersecurity (ENISA) is a centre of expertise for cyber security in Europe and is located in Athens, with a second office in Heraklion [88]. The country aims to become the heart of the European Union (EU) cybersecurity. Since 2019, Greece has ranked first among 160 countries in the National Cybersecurity Index (NCSI). This measures the national cyber security capacity building and preparedness of countries to prevent cyber threats and manage cyber incidents [89]. This ranking implies that Greece is continuously seeking to ensure cyber safety and develop secure websites for its users [16]. The Greek National Cybersecurity Strategy is also considered as one of the most comprehensive across the European Union [17]. The country is also currently 28th on Global Cybersecurity Index and 38th on the Information and Communication Technologies (ICT) Development Index. All these show that the development and commitment towards cybersecurity have been strongly promoted in Greece since 2019 [16]. Perhaps due to Greece's cybersecurity strategy, Facebook has opted to use it to trial novel methods to mitigate deepfakes and misinformation online.

Finally, Greece is a worthy case to investigate further, as for the past few years the profile of the country has high levels of untrust towards the Internet and social media [90]. A recent study has shown that the Net Trust Index of Greece is -2, which puts it into the category of Low Trust and continues to slowly decline. The Net Trust of Greek citizens towards social media networks is in the lowest category of No trust and has recorded a significant decrease over time [90]. Most Internet users in Greece (72.5%) seem to be sceptical regarding the reliability of the information they access online. They also perceive themselves as rather capable to distinguish fake news and images on the Internet [19].

The ability of Greeks in detecting fake content has been leveraged by Facebook. However, no study challenging their abilities to detect deepfakes online has been conducted on Greek users. At the same time, several studies investigate the perceptions of the Greek users towards the social media platforms and Internet. There is no research on the exposure of Greeks to deepfakes online, their perceptions towards the deepfakes, and their cybersecurity. Also, this chapter introduced a range of deepfakes detection methods based on different techniques. It is worth exploring the methods used in Greece for detecting deepfakes. The novel approaches of Facebook and Ellinika Hoaxes are taken under investigation to provide some insight into the effectiveness of these methods in the detection of deepfakes.

To summarise the chapter, deepfakes is a well-defined and analysed term in literature and the main features of deepfakes have been discussed and analysed. The deepfakes spectrum varies greatly and includes both cheaply manipulated content but also highly sophisticated AI-generated audio-visual content. Since, their occurrence, they have been mostly used for entertainment as well as malicious purposes. Researchers have noticed that the use of deepfakes can severely damage to individuals, organisations, and states. This concern is addressed by the efforts to develop tools which aim to detect all kinds of deepfakes with the highest efficiency and accuracy possible. These actions aim to prepare organisations and such as social media platforms to protect their users from malicious manipulations of online audio-visual content. The human ability to detect deepfakes can be combined with the deepfakes detection technology. Tests conducted show that humans can detect deepfakes which machine learning and other detection methods are not. This combination could provide the best results in deepfakes detection.

# 4. Results

This chapter presents the data obtained by collecting responses from Greek social media and Internet users to an online survey. Annex B features the questionnaire shared. It was distributed via email and Facebook to **150** individuals and received a total of **123** replies, collected within the period 8th October 2021 till 24th November 2021. All participants were made aware of the topic and purpose of the questionnaire. Anonymity was ensured as no identifiable personal information, such as full name and email was requested. This safeguards privacy of the responders and enhance the honesty and accuracy of their responses. No ethical issues were raised within this research.

The analysis of the data seeks to describe the deepfakes landscape in Greek social media and the Internet. It explores the trends, attitudes, and preferences of Greeks in social media platforms and website content and rates the exposure of Greeks to online deepfakes. In addition, this chapter details the accuracy of deepfakes detection by those who completed the questionnaire. Finally, the chapter provides an overview of users' perceptions towards their own cybersecurity and knowledge of deepfakes. These results rely upon the analysis of collected responses rated in the 5-point Linkert scale from 1 as Strongly Disagree to 5 as Strongly Agree.

## 4.1 Detection of deepfakes by Greek social media and Internet users

The rise of the threat of online deepfakes brought the detection of manipulated content to the notice of governments and the security industry. While machine learning models and technical tools are continuously being researched and developed, human detection capabilities are less explored. Some experts note that the most effective detection of malicious deepfakes requires a combination of deepfakes detection technology and human perception [81] [82] [83]. From studies on the human visual system, it has been shown that it uses mechanisms for facial recognition and utilises a dedicated region of human brain focused on face perceptions. Exploiting this natural ability could be used to enhance the deepfake detection performance of humans [113].

A survey from 2020 showed that 72.8% of Greeks participants were capable of distinguishing fake news and content online [19]. The deepfakes detection test conducted within this thesis takes into consideration this percentage. It challenges the hypothesis that the participants will score a detection accuracy higher than 72.8%.

The questionnaire included 10 images, of which 6 were deepfakes and remaining 4 were authentic images. It also presented 3 videos, 2 artificially manipulated and one authentic video. The maximum score which could be achieved was 13 for both images and videos. The images of the questionnaire can be viewed in the Annex D of this thesis.

Only 3 responders correctly characterised all images and videos included in the questionnaire. This corresponds to 3.69% of all participants. The average (in terms of "mean") person selected correctly 9 out of 13 images or videos. The overall accuracy for images detection stands at **69,3**%. The average person had 6 images detected correctly. For the videos the average score is 3. This matches with 64,1%% - for images - and 87,0% -for videos - accuracy respectively.

### 4.1.1 Images

Analysis of the survey responses discovered the most common mistakes among the participants. The lowest detection accuracy is 31.7% for a cheap deepfake image, as seen in the Figure 15. Except for the first deepfake image, the other deepfakes which were mistaken for authentic images were manipulated with the use of FaceApp. This application offers a variety of manipulation tools and editing effects on images. These deepfakes are considered "cheap" deepfakes. They were created with publicly available software and do not require an expertise in machine learning and sophisticated software to be created.

The 1st image was a deepfake, which was detected by only 39.8% of participants, as seen in the Figure 15. This constitutes the second lowest accuracy. The image is the only highly sophisticated deepfake included in the questionnaire. It was downloaded from https://www.whichfaceisreal.com/index.php. Based on the findings, it becomes clear that the participants were fooled and did not identify the image as AI generated.

The third highest misidentification of a deepfake collected 43 correct responses which results in 35% accuracy of all participants. This image was a cheap fake created with FaceApp. Based on above, the lowest detection accuracy was of a cheap fake. People can be easily deceived by cheap fakes. The results indicate that cheap fakes can be as effective as sophisticated deepfakes. 30-40% of users were confused by both cheap and highly sophisticated deepfakes.

Another image that was not correctly identified by the majority of the responders was the 10th image of the questionnaire, which is authentic. Specifically, only 49 out of 123 responders opted for authentic image, which corresponds to 39.8% of all participants. Table 5 displays the percentage of correct responses and mishits.

The remaining 6 images were accurately identified by most of the participants. Specifically, the 2nd image in the questionnaire is a deepfake manipulated with the use of FaceApp. 85.4% of responders were successful in its identification as shown by Figure 15 below.

The 3rd image was picked as authentic by almost 75% of participants as shown in Figure 15. The 5th and 6th images, both authentic, gathered more than 80% correct answers. The 7th image is a deepfake and more than 74% of responders made the correct decision to mark it as deepfake. Finally, the highest accuracy is also visible in the following Figure 15. More than 90% of participants marked the image as a deepfake, which corresponds to 113 correct answers.

Overall, the highest score in images detection was 100% and the lowest was 30.76% (4 correct out of 13). The overall detection accuracy was 74.89%. All the accuracy results are gathered in the following Table 5. The 4th image collected the lowest detection accuracy (31.7%), while the 9th image was detected by 91.9% of the survey participants. Users scored highest accuracies for authentic images compared to deepfakes. Thus, it was easier for the users to be deceived by deepfakes, while cheap fakes (Image 2, 4, 7, 8, and 9) were more confusing for them compared to the highly sophisticated deepfake (Image 1).

Table 5. Correct and wrong image detection in percentages

| Image | Correct answer | Wrong answer |
|---|---|---|
| **Image 1: Deepfake** | 39.8% | 60.2% |
| **Image 2: Deepfake** | 85.4% | 14.6% |
| **Image 3: Authentic** | 74.8% | 25.2% |
| **Image 4: Deepfake** | 31.7% | 68.3% |
| **Image 5: Authentic** | 86.2% | 13.8% |
| **Image 6: Authentic** | 81.1% | 17.9% |
| **Image 7: Deepfake** | 74.8% | 25.2% |
| **Image 8: Deepfake** | 35.0% | 65.0% |
| **Image 9: Deepfake** | 91.9% | 8.1% |
| **Image 10: Authentic** | 39.8% | 60.2% |

This table shows the percentages of correct and wrong answers per image. The following Figure 15 shows graphically the accuracy percentages.

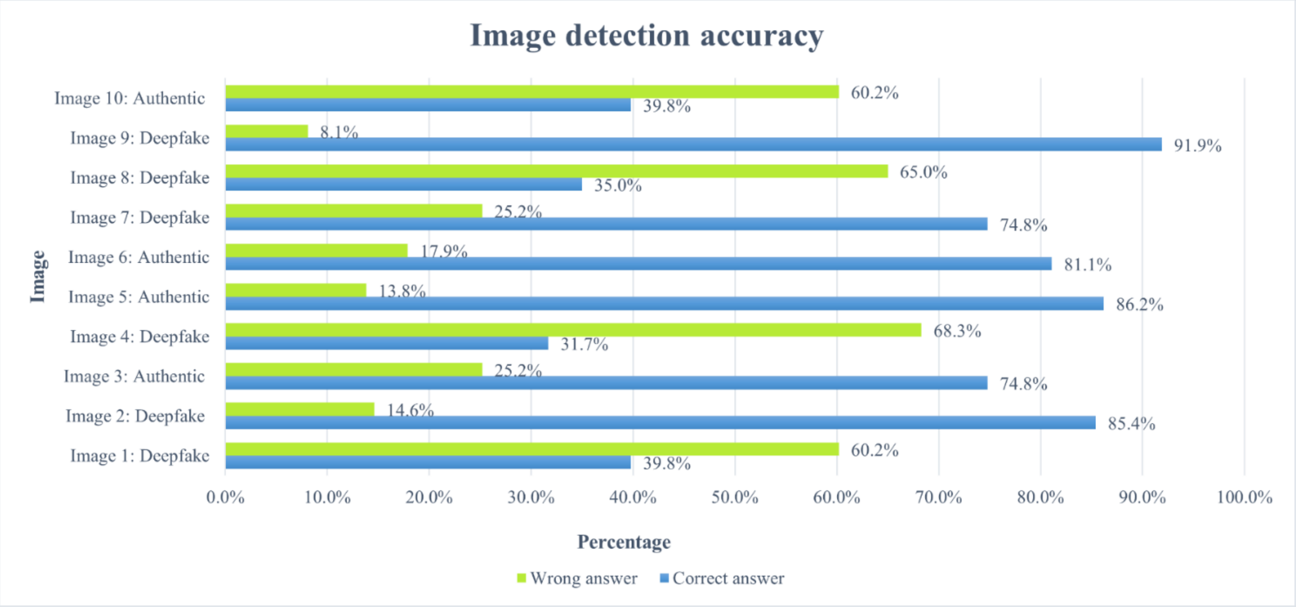


Figure 15. Detection percentages per image

### 4.1.2 Videos

The participants were presented with 3 videos, 2 of which were deepfakes, generated with the use of the 'First Order Model' [93]. A view of the accuracy percentages in the graphs below illustrates that most participants detected correctly the artificially manipulated content.

The 1st video was detected as deepfake by 108 out of 123 participants, which is an accuracy of 87.7%. This can be viewed in the Figure 16. The 2nd video is also a deepfake with 100 people correctly selecting it as artificially manipulated content (81.3%) with only 23 believing the video was authentic, as illustrated in the Figure 16. Finally, the 3rd video was authentic and recorded the highest accuracy at 91.1%, which means that only 11 or 8.9% individuals chose the wrong answer. The results can also be viewed graphically in the Figure 16. The average percentage for the three videos is 86.76%. This illustrates the comparative ease of producing high quality fake images and the challenge of creating convincing vide material. The motions and body language, the eye movements and facial expressions could be significant elements which may help in detecting artificially manipulated or generated content.

All in all, the following Table 6 describes in a key-coded manner the percentages of correct and wrong answers for each video in the questionnaire. The 3rd video was detected by 91.9% of the survey participants, which is the highest detection accuracy listed in the Table 6.

Table 6. Video detection percentages

| **Video** | **Percentage of correct answers** | **Percentage of wrong answers** |
|---|---:|---:|
| **Video 1: Deepfake** | 87.80% | 12.2% |
| **Video 2: Deepfake** | 81.3% | 18.7% |
| **Video 3: Authentic** | 91.9% | 8.9% |

The Figure 16 illustrates this data in a form of graph, depicting the percentages for wrong answers in yellow colour and for correct answers in blue colours.

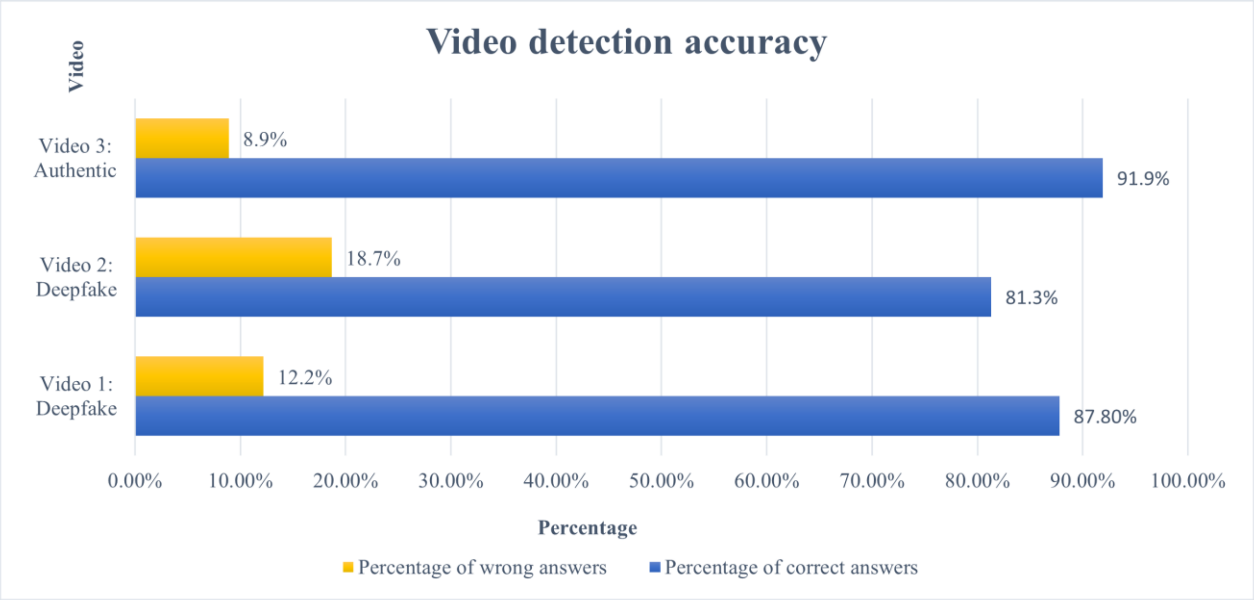


Figure 16. Detection percentages per video

All in all, the Table 7 shows the accuracies scored per content. The Table 8 describes the overall accuracies for all images, for all videos separately and all content collectively.

Table 7. Detection accuracies per content

| **Content** | **Accuracy** |
|---|---|
| **Image 1** | 39,8% |
| **Image 2** | 85,4% |
| **Image 3** | 74,8% |
| **Image 4** | 31,7% |
| **Image 5** | 86,2% |
| **Image 6** | 81,1% |
| **Image 7** | 74,8% |
| **Image 8** | 35,0% |
| **Image 9** | 91,9% |

| | |
|---|---|
| **Image 10** | 39,8% |
| **Video 1** | 91,9% |
| **Video 2** | 81,3% |
| **Video 3** | 87,8% |

Table 8. Average detection accuracy per content

| | | |
|---|---|---|
| **Average accuracy** | **All content** | 69,3% |
| **Average accuracy** | **Images** | 64,1% |
| **Average accuracy** | **Videos** | 87,0% |

The survey participants scored higher accuracy in detecting videos, while their results were significantly lower in detecting images. In general, the accuracy for all content included in the questionnaire, they reached 64,1% of detection accuracy.

## 4.2 Social media and Internet usage

Social media and the Internet play an important role in the proliferation of deepfakes as they are the catalyst for them to go viral.  To examine this issue, the second section of the questionnaire focused on attitudes and preferences of Greeks towards the social media they use and the content they search for online. It also explored the exposure of Greeks to deepfakes on those platforms. The users were asked a series of closed questions regarding social media, Internet use and deepfakes. All questions required an answer with some allowing multiple choices while others only permitted a single response. The results indicated which social networks and websites should gain more attention from policy and law makers for the purpose of deepfakes regulation. It is important to note that the questionnaire includes the following statement:

*"A deepfake is a photo, video, or audio track created using artificial intelligence techniques to realistically simulate or alter people's faces, movements, and voices,*

*among other simulations. Please give your opinion whether the following images and videos are deepfake or real."* The purpose of this statement is to clarify what constitutes a deepfake, as many users may not be aware of the term and the fact that the common content manipulated with Instagram, Facebook, etc. filter is a deepfake.

### 4.2.1 Social media

The users were asked which social media platforms they use. The social media chosen as options for these questions were the most popular platforms in Greece [114] [115] [116]. All these platforms already have cases of deepfakes or have the potential of deepfakes occurrence [117]. According to the responders, the most popular social media applications are Facebook (93.5%), YouTube (92.7%) and Instagram (78.9%). Reddit (9.8%), TikTok (23.6%) and Twitter (32.5%) are the least used social media platforms. All results are visible in the Figure 17.

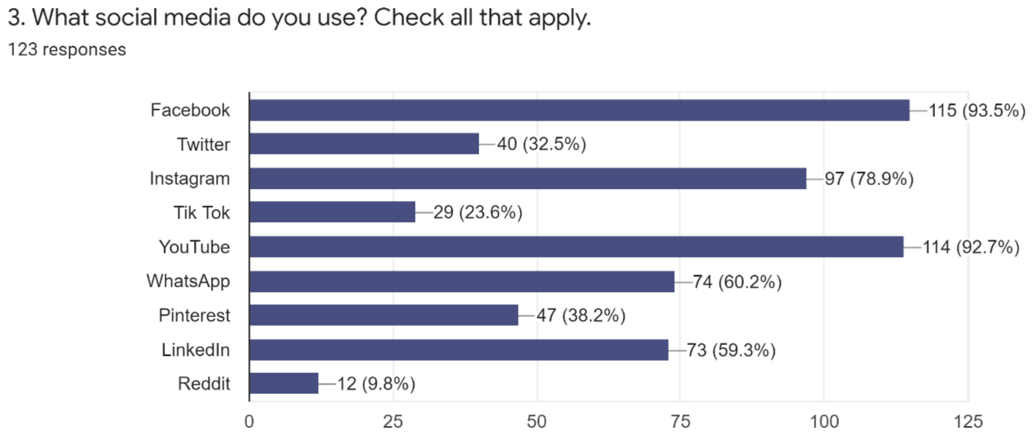


Figure 17. Social media popularity

The participants were also asked to measure the time spent on social media. The possible options were measured in the scale of hours on a daily basis. The results can be seen in Figure 18.

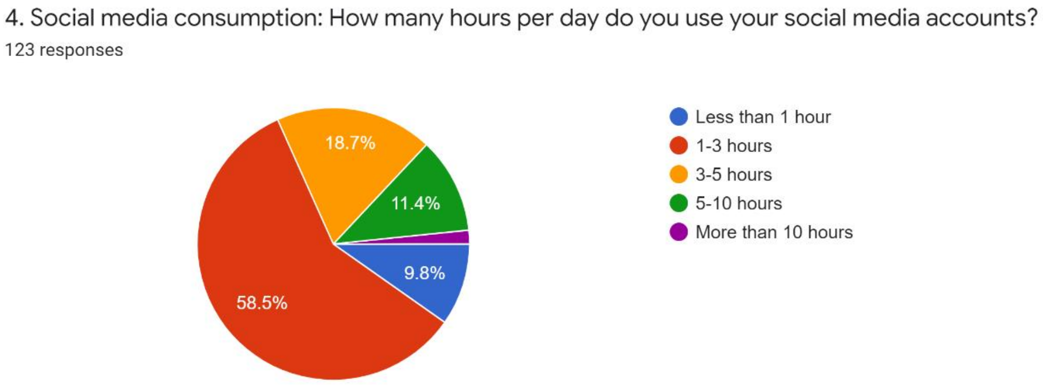


Figure 18. Social media daily consumption

Most of the responders spend approximately 1-3 hours per day on social media platforms. Almost 20% spend 3-5 hours daily using their social media accounts. 11% of responders allocate 5-10 hours during the day on social media, while around 10% uses social media accounts less than 1 hour per day.

The questionnaire asked users to select the social media platforms with most deepfakes viewed. According to the responders, the most deepfakes were seen on Facebook (63.4%), Instagram (65%), and YouTube (45.5%). This is expected as those platforms gain the highest popularity among responders. TikTok (21.1%) and Twitter (11.4%) follow. Those are the least preferred social networks. Despite this, the users noticed that those platforms contain a significant number of manipulated images and video. of deepfakes. A full image of the responses can be viewed in Figure 19.

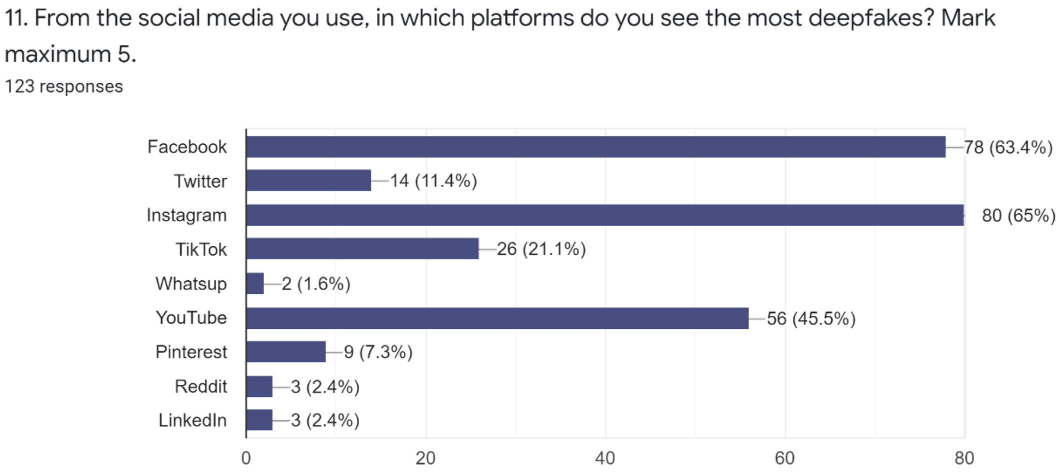


Figure 19. Social media with most deepfakes

Additionally, each participant was asked to provide an estimation of the number of deepfakes encountered daily on social media platforms. The questionnaire clarified what constitutes a deepfake, so that users would become aware that content manipulated with Instagram, Facebook filters, or photoshopped are examples of deepfakes. However, the metrics are based upon each user's understanding of what is a deepfake. Specifically, most users estimate that they 2-5 deepfakes on social media. 20% of them view from 6 to 10 deepfakes, while only 3 participants responded that they view 10-15 deepfakes (2.4%). 7 individuals considered that they experience every day more than 15 deepfakes while using their social media accounts. This corresponds to 5.7% of all participants. Finally, 7.3% of all users responded that the do not see any deepfakes on social media. This equals to 9 out of 123 individuals. All above are presented graphically in Figure 20.

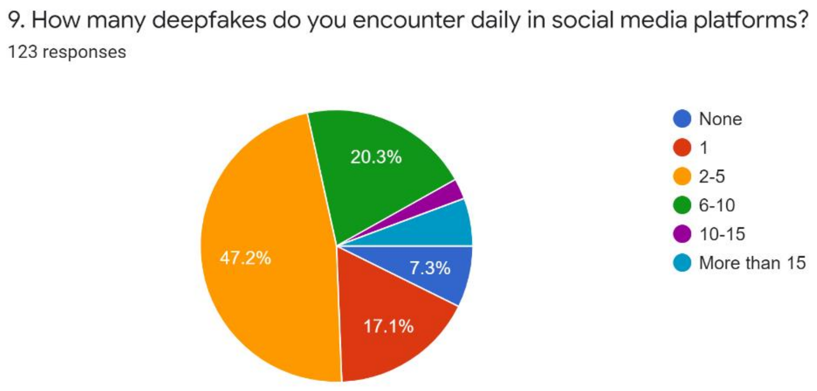


Figure 20. Estimated daily number of deepfakes viewed on social media

### 4.2.2 Internet

The questionnaire explored the attitudes and preferences of Greeks on the wider Internet and not only on social media. The questionnaire sheds light on the most popular website categories that users visit. All responses, alongside the numbers and percentages can be viewed in the Figure 21, as below.

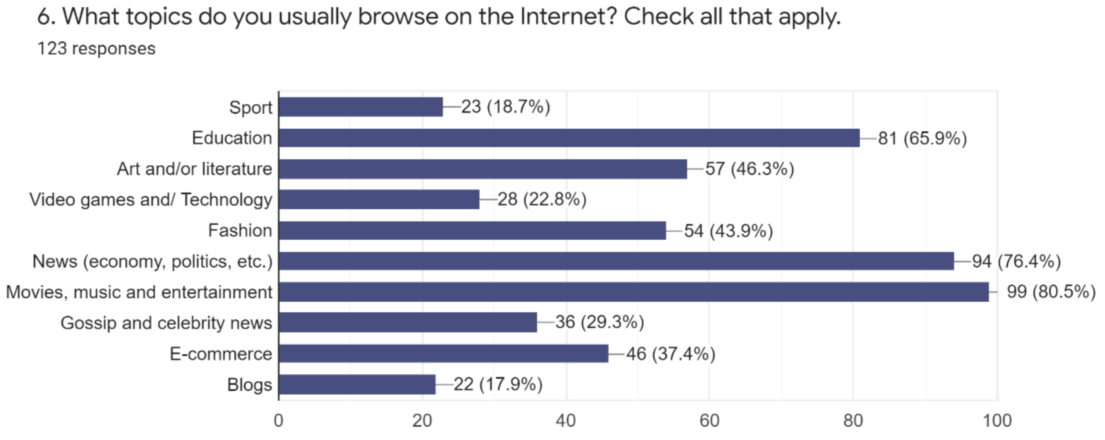


Figure 21. Popularity of websites in Greek Internet

The top three spots belong to movies, music, and entertainment (80.5%), news (76.4%) and education (65.9%). The least popular websites are related to blogs (17.9%), sport (18.7%) and video games and/or technology (22.8%).

Regarding social media, users were asked to measure their time spent daily on the Internet. The results are presented in the Figure 22. Most users recorded 1-3 hours per day, but the proportion of users who spend 3-5 hours daily are not significantly behind. While 38.2% of responders consume 1-3hours browsing the Internet daily, those who surpass 3 hours but do not exceed 5 hours stand close at 26.8%. 18 participants responded to use Internet less than 1 hour per day, while users who exceed 5 hours are 25 out of 123, which corresponds to 12.2.% and 8.1% respectively.

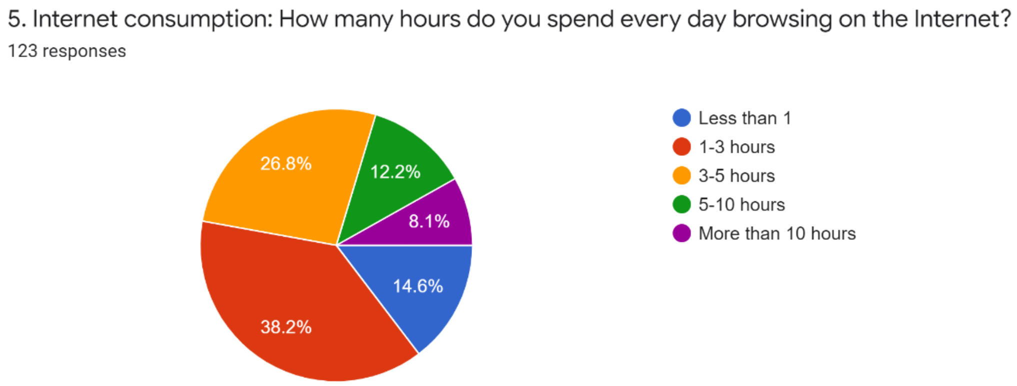


Figure 22. Daily Internet consumption

The responders were requested to choose the website categories with the estimated highest proportion of deepfakes. The results are shown in Figure 23. It is not surprising that movies, music, and entertainment attract a high number of responses as it is among the most popular topics. At the same time, this is also expected as deepfakes and AI are commonly used in this industry.

The highest response is the category of Gossip and Celebrity news. This also confirms the current trend that deepfakes target known personalities and celebrities. Education, Art and/or History as well as E-commerce have the least number deepfakes, according to Greek users. Figure 36 illustrates all the collected responses.

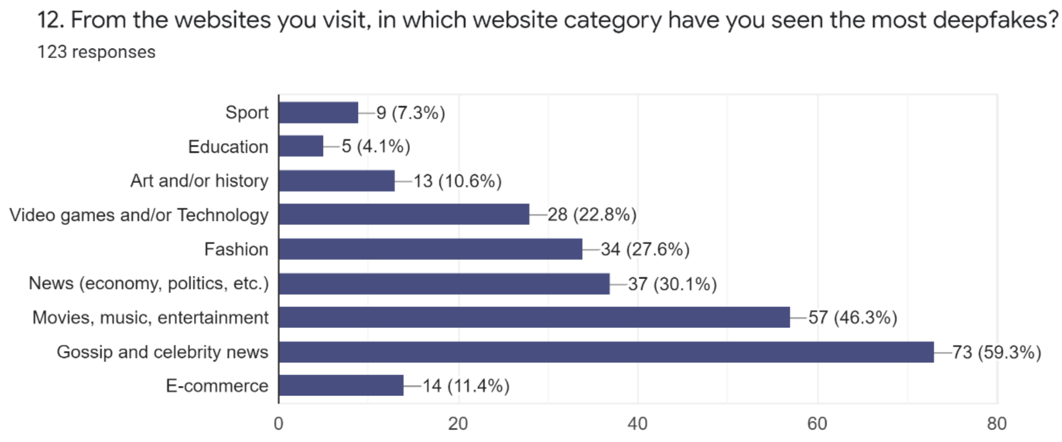


Figure 23. Websites with most deepfakes

Finally, on the question "How many deepfakes do you encounter daily while browsing on the Internet", the greatest response answer is 2-5 deepfakes. More than 46% of participants (57 out of 123 individuals) stated that they viewed this amount of deepfakes. The second most recorded answer belongs to "None". 20 responders claim to see no deepfakes while browsing the Internet. The responses of "1 deepfake" and "5-10 deepfakes" hit similar percentages (13% and 14.6% respectively). More than 10 or 15 deepfakes are the least common answers among the responders (6 responses for each). All the results are represented graphically in the Figure 24.

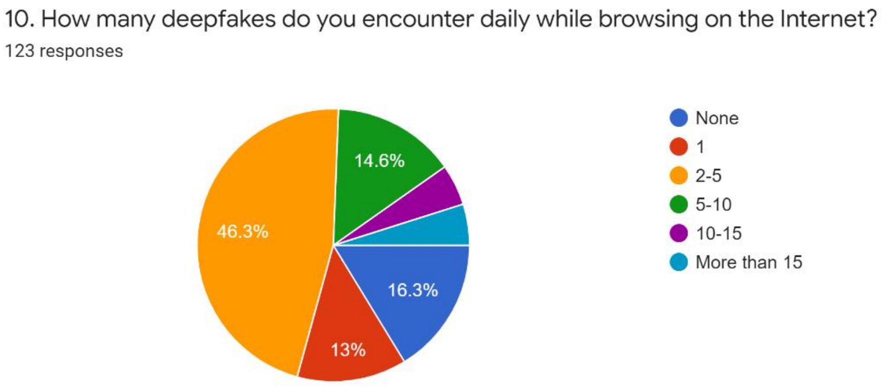


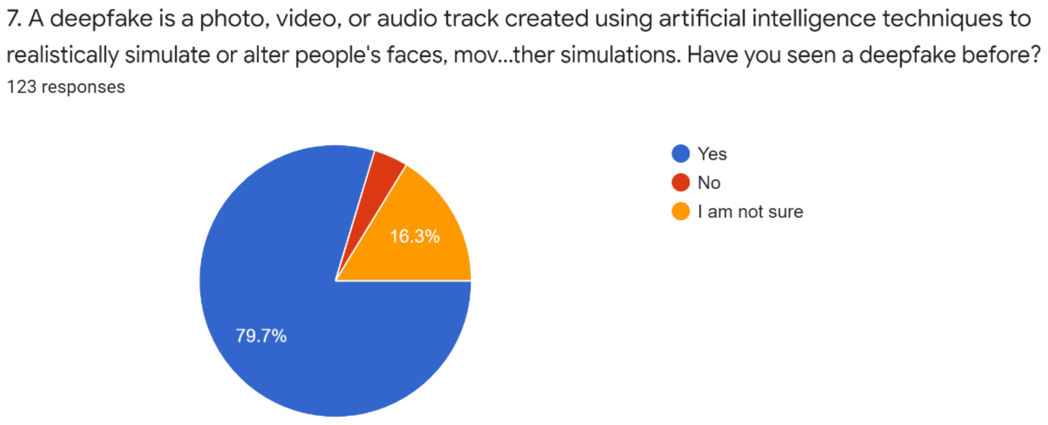
Figure 24. Daily amount of deepfakes on Internet

## 4.3 Perceptions towards deepfakes

The perceptions and experiences of Greek online and social media users are at the heart of the third section of the questionnaire. Each participant was requested to rate the level of their agreement with some statements regarding deepfakes. The questionnaire used the five-point Likert scale ranging from “strongly disagree” to “strongly agree” (1-Strongly Disagree, 2 - Disagree, 3 – Neutral, 4 – Agree, 5 - Strongly Agree).

Before this, the second section of the questionnaire addressed some general questions on deepfakes. They aimed to depict if users understand the term “deepfake” and can name categories of industries with deepfakes occurrence. The users were also asked to express their feelings associated with deepfakes and highlight the strongest one.

First question which was addressed to the participants was “Have you seen a deepfake before?”. The Figure 25 illustrates the portions of each answer.



7. A deepfake is a photo, video, or audio track created using artificial intelligence techniques to realistically simulate or alter people's faces, mov...ther simulations. Have you seen a deepfake before?

123 responses

Figure 25. Prior experience with deepfakes

Only 4 individuals claimed that they have not seen deepfakes before. This equals to 4.1% of responders.  98 out of 123 (almost 80%) believe they have seen deepfakes. Finally, 20 responders (16.3%) were not fully sure they have seen AI manipulated or generated content.

Supporting the above, the users were asked to name sectors in which they have encountered a deepfake. Figure 26 illustrates the responses collected. The top responses are social media filters on apps (74.5%), photoshopped, cropped or manipulated content

(71.5%), memes (71.5%), and satire videos of politicians or celebrities (61%). Art and/or history as well as health collected the least responses, 15 and 8 out of 123 respectively. This may come from the fact that in Greece those sectors are not highly digitalised and innovative.

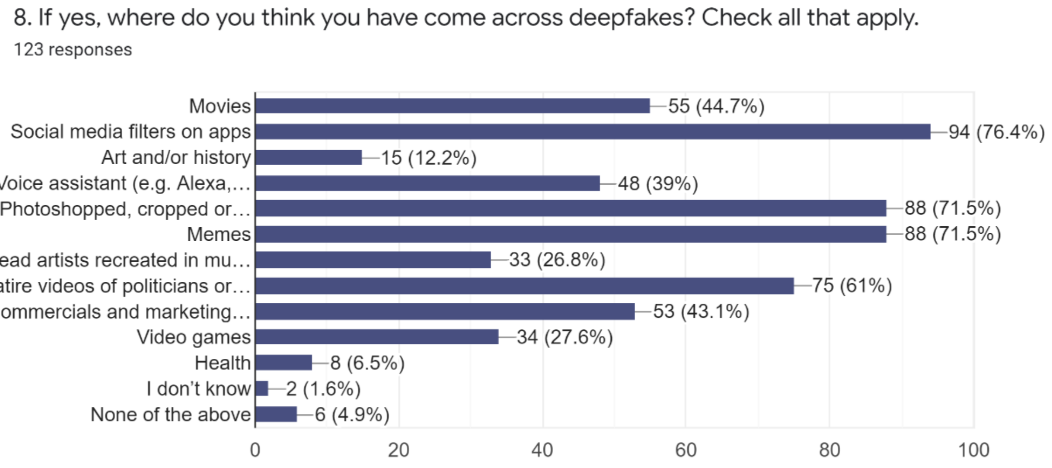


Figure 26. Sectors with most deepfakes occurrence

Additionally, the users were asked to express their feelings associated with deepfakes.

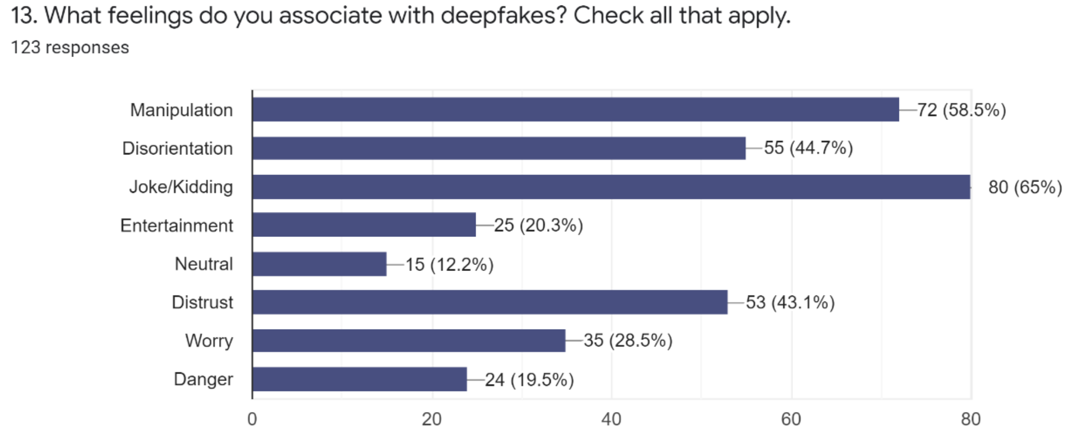


Figure 27. Feelings towards deepfakes

The results presented in the above Figure 27 show that for most users deepfakes do not have a negative correlation. The main feeling "Joke/Kidding" was perceived by 65% of responders. Deepfakes used to distort reality and delude targeted audiences was expressed by 58.5% of participants, who feel manipulated due to deepfakes. 55

individuals claimed to feel disorientation and only a few less noted that deepfakes create a feeling of distrust. Very few responders express feelings of danger associated with deepfakes.

Figure 28 shows the strongest feeling associated with deepfakes. Manipulation was gathered the most responses, followed by feelings of distrust and disorientation. Overall, those results ascertain the opinion that deepfakes are the new form of digital manipulation [118] [119] [120].

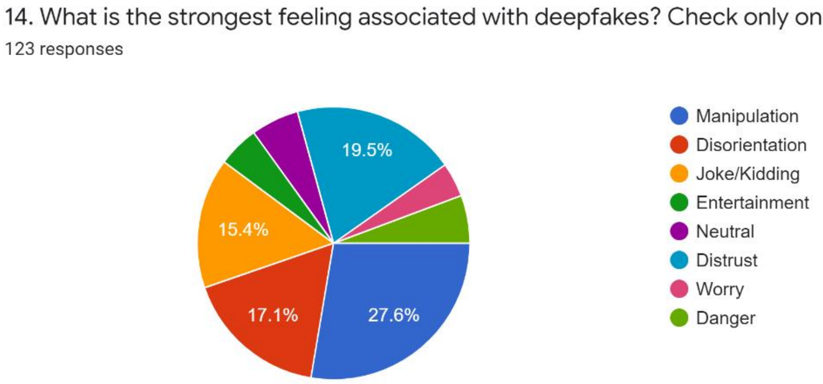
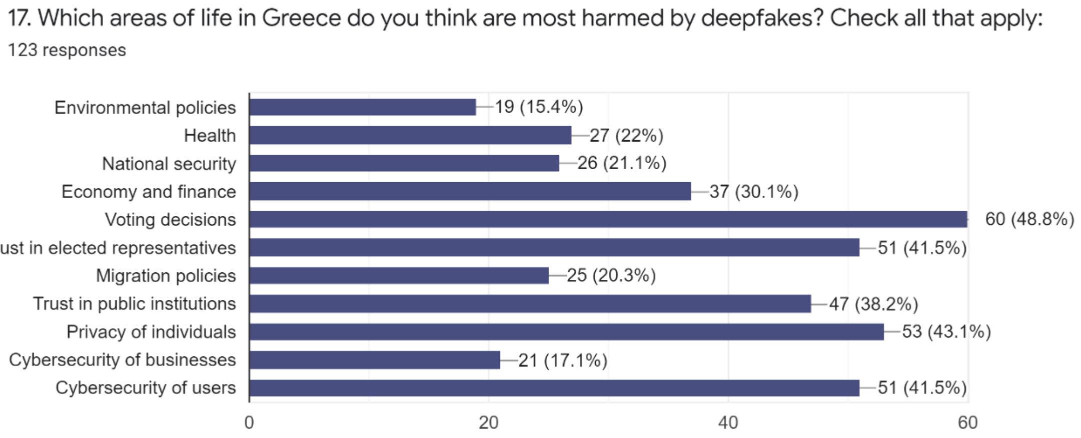

14. What is the strongest feeling associated with deepfakes? Check only one.
123 responses

Figure 28. Prevailing feeling towards deepfakes

Finally, the questionnaire addressed the issue of the implication of deepfakes on the state, economy, and society, as seen in Figure 29. Specifically, most responders believe that deepfakes have caused the biggest damage to voting decisions in Greece. Another implication which is highly expressed by the participants is the loss of their privacy.

Finally, trust towards elected representatives as well as the cybersecurity of users are also significant areas negatively impacted by deepfakes.



17. Which areas of life in Greece do you think are most harmed by deepfakes? Check all that apply:

123 responses

Figure 29. Sectors harmed by deepfakes

The third section of the questionnaire gave a detailed insight into the experiences of Greeks with deepfakes. The participants were asked to rate their level of agreement with the statements addressed to them.

A detailed overview of all statements and results collected from the users' responses can be accessed in the Annex B.

The key elements of the results are the following:

- Most users do not agree that deepfakes in Greek social media and Internet are malicious.
- Most users have not experienced a cyberthreat with the use of deepfake, such fraud and phishing.
- Most users do not feel confident in the current efforts and actions taken from the Greek government and legislation bodies, the social media companies, and educational and awareness institutions to combat deepfakes.
- Most users do not agree that Greek government holds the main responsibility to fight the malicious deepfakes online.

- Most users are confident that a deepfakes national strategy should be put into place to address to issue of malicious deepfakes and raise their cybersecurity.

# 5. Analysis

The first section of the questionnaire tested the users' ability to detect manipulated content. The results obtained are shown in the following Table 9. Those results are also the dataset used for conducting hypothesis testing and significance analysis.

Table 9. Deepfakes detection accuracy per content

| **Content** | **Accuracy** |
|---|---|
| **Image 1: Deepfake** | 39,8% |
| **Image 2: Deepfake** | 85,4% |
| **Image 3: Authentic** | 74,8% |
| **Image 4: Deepfake** | 31,7% |
| **Image 5: Authentic** | 86,2% |
| **Image 6: Authentic** | 81,1% |
| **Image 7: Deepfake** | 74,8% |
| **Image 8: Deepfake** | 35,0% |
| **Image 9: Deepfake** | 91,9% |
| **Image 10: Authentic** | 39,8% |
| **Video 1: Deepfake** | 91,9% |
| **Video 2: Deepfake** | 81,3% |
| **Video 3: Authentic** | 87,8% |

The main assumption of the thesis is that the participants will score higher detection accuracy than 72,8%. The Table 10 describes the average detection percentages per content.

Table 10. Average detection accuracy per content

| **Content** | **Average detection accuracy** |
|---|---|
| Images | 64,1% |

| | |
|---|---|
| Videos | 87,0% |
| Images + Videos | **69,3%** |

The null hypothesis ($H_0$) is that the average detection accuracy (d) will be equal or less than 72.8%. The alternative hypothesis ($H_A$) is that the detection accuracy (d) is greater than this percentage.

$H_{0:}$ d ≤ 72.8%

$H_{A:}$ d > 72.8%

To test the above and estimate the statistical significance, a t-test is conducted, and two numbers are calculated. Specifically, the p-value and a cut-off value, α. The p-value represents the smallest value of α for which the null hypothesis can be rejected [121]. Regarding the cut-off point, 0.05 has been widely used as cut-off for statistical significance [122] [123].

The following Table 11 shows the results of the t-test conducted on the sample. Three significant digits have been maintained for all calculations for more clarity and readability in terms of understanding and comparing the The database consisted of all images and videos detection accuracies measured in percentages. The cut off value chosen is 0.050. The p value is **0,302**. As it is larger than the cut-off value α, this means that the null hypothesis $H_0$ is failed to be rejected with α=0.05. The alternative hypothesis $H_A$ is not accepted. A "failure to reject" a hypothesis should not be confused with acceptance [124].

The calculated p-value indicates that the evidence favours the null hypothesis, and the results are considered as not statistically significant.

Table 11. t-Test on deepfakes dataset

| **t-Test** | | |
|---|---|---|
| | | |
| | *Accuracy* | |
| Mean | 0,693 | |
| Variance | 0,054 | |
| Observations | 13 | |

| | | |
|---|---:|---|
| Hypothesised Mean Difference | 0,728 | |
| df | 12 | |
| t Stat | -0,531 | |
| **P(T<=t) one-tail** | **0,302** | |
| t Critical one-tail | 1,782 | |
| P(T<=t) two-tail | 0,604 | |
| t Critical two-tail | 2,178 | |

According to Frank Slide such result could mean that the findings are not strong enough to suggest that an effect exists in the sample population [125]. At the same time, the author insists that the effect could exist, but it can be minor, or the sample size is not big enough.

All in all, the analysis performed on the dataset of collected accuracy results does not support the key assumption of the thesis. Specifically, it was assumed that Greek social media and Internet users have high levels of detection accuracy of at least 72.8%. The participants of the questionnaire did not reach this target. It was estimated that the detection accuracy was 69.3%. This chapter dealt with the RQ2 “Can Greek Internet and social media users tell apart deepfake videos and images from real ones?”. It is answered negatively that Greek Internet and social media users can tell apart deepfake videos and images from real ones.

# 6. Discussion

Greece has not incorporated the topic of deepfakes into its legislative system and regulations. The legal tools that could assist in deepfakes regulation, such as copyright law, AI law and cyber crime law are outdated. They also do not refer to recent issues that derive from technological advancements and AI. The legal landscape on deepfakes is more developed in the EU where law and regulations follow and respond to some of the latest technological developments. These seem more relevant and fitting to the current problems rising from new technologies, AI, and digitalisation. It is worth mentioning that in both cases the laws on AI are still in the process of being passed and implemented. This demonstrates that developments are more advanced within the EU rather than in Greece with the former showing more confidence and decisiveness with its law proposals and plans. At present EU law seems to be the main platform for regulating deepfakes, while Greek law is complimentary in some elements. In answering **RQ1** “What is the current legislation in Greece on deepfakes?”, it was assumed that there is no legislation in Greece regulating deepfakes technology. **This assumption is correct, as no law directly regulates the creation and use of deepfakes in Greece**. However, specific articles on law could indirectly be used to in legal battles involving the malicious use of deepfakes.

The **RQ2** investigated the detection capabilities of a sample of the Greek population. In general, responders scored an overall deepfakes detection accuracy of 69.3%. They were challenged with 13 pieces of content. Specifically, they were presented with 6 deepfake images and 2 deepfake videos and 4 real images and 1 authentic video. A survey from 2020 showed that 72.8% of Greeks participants were capable of distinguishing fake news and content online [19]. The test conducted within this thesis challenged the assumption that Greek social media and Internet users will score higher detection accuracy for deepfakes. The test showed that this percentage was not met. A reason for this could be those technological advancements in deepfakes production makes them more difficult to detect. This may highlight the increasing threat posed by malicious deepfakes in the future.

The assumed target accuracy of deepfakes detections was 72.8%. The results of the questionnaire show that the participants do not have this effectiveness of the detection.

In short, the answer to the **RQ2** “Can Greek Internet and social media users distinguish deepfake videos and images from real ones?”, **has not been proven by the data**. The overall detection accuracy of videos is significantly higher than the detection of images. Specifically, the average percentage for detecting videos was 87%, while for images was 64.1%. The results show that it was easier to detect deepfake videos compared to images.

Videos incorporate facial expressions, feature movements and body language which could manifest artificially manipulated content. Those dynamic elements can more easily show some irregularities, unusual positions, misalignments and incongruencies in different dimensions and movements than the static elements of an image.

Additionally, some elements which could indicate a deepfake are the blinking of the eyes, unnatural lighting and shadows near the eyes, eyebrows and face or body angles. Facial hair which does not look real and natural, lack of wrinkles or unusual skin movements are also examples of deepfakes indicators. Another sign of a deepfake could be the mouth looking disconnected from the rest of the face, or the head from the rest body. Lack of emotions and facial expression could also constitute a deepfake flag.

As most deepfakes spread online show famous personalities, reverse image search could be a useful method to detect a deepfake. The questionnaire showed that most Greek participants are neutral towards this method. This could mean that they do not know what reverse image search is.

**RQ3** investigated “What is the impact of malicious deep fakes on the Greek Internet and social media users, based on their perceptions and experiences in the cyber domain?”. The findings of the questionnaire suggest that most Greek social media and Internet users claim to have seen a deepfake. The most common means propagating deepfakes are filters on social media apps. Memes and photoshopped, cropped or manipulated content are also very common deepfakes examples.

Social media platforms are a medium for the spread of deepfakes. The most popular social media platforms in Greece are Facebook, Instagram, and YouTube, followed by WhatsApp and LinkedIn. The users of social media in Greece believe that they encounter an amount of 2-5 deepfakes daily on several social media platforms, during a daily consumption of 1-3 hours. According to the interviewed Greeks, most deepfakes are circulated on Facebook, Instagram, YouTube, TikTok and Twitter. Thus, the detection and removal of deepfakes in Greece should start and focus on the above social media platforms.

As well as social media, Internet websites also contribute to the dissemination of malicious deepfakes online. The most popular website categories searched by the interviewed users are film, music and topics related to entertainment, news such as politics and economy and education content.

While browsing the Internet, most users estimate that they encounter 2-5 deepfakes per day of Internet surfing during a daily Internet consumption of 1-3 hours. Gossip and celebrity news as well as films, music and entertainment topics are the top three website categories with most deepfakes. The above results align with the reported trend that most deepfakes are portraying celebrities and famous personalities and in music and film industries [126] [23] [60] [54] [127] [128].

For most users, deepfakes are associated with feelings of amusement and joking. Thus, they do not currently view deepfakes as malicious. However, some responders do claim to experience feelings of manipulation and disorientation as deepfakes distort reality. However, there is a feeling that deepfakes will pose a threat in the future. The greatest threat posed by malicious deepfakes are voting decisions in Greece, cybersecurity and online privacy of Greek users and their trust towards the elected representatives and government. These areas are often described by the experts as the greatest threat posed by malicious deepfakes [60] [14].

Deepfakes could become a means to influence elections due to their ease of production and ability to rapidly spread fake narratives and misinformation, creating fictitious content against governments and politicians. Deepfakes can distort the democratic discourse, erode trust in public institutions, incite political divisions and violence and destabilise political situation and processes within a country. Greece has suffered from political instability over the past decade with its democratic system being challenged by polarised and extreme politics [129] [130]. Greece will have elections in 2023 and misinformation campaigns with the use of malicious deepfakes should not be excluded as a potential weapon of the political parties.

At the same time, it was wisely noted by the questionnaire participants that deepfakes can potentially pose a threat to their online privacy and cybersecurity in the future. Criminals are capitalising on the dark side of deepfakes technology to commit online frauds, steal information and identities. The general approach to the content posted online and on social media is lack of trust from the users and a critical approach to the material they see. At the same time, most Greek participants do not find themselves confident in detecting AI-manipulated content. Based on the experiences of the questionnaire participants it cannot be clearly determined if the content posted online and on social media is misleading or harmful for the users. Most Greeks tend to believe that deepfakes are not malicious and can benefit several spheres of human activities such as education and entertainment.

Experience of phishing attempts and fraud with the use of malicious deepfakes was not reported often by the questionnaire participants. The threat of deepfakes in Greece is not commonplace, but according to the users more malicious deepfakes will target users and undermine their cybersecurity in the future. This is particularly so when using social media or browsing the Internet. They also expressed worries over their cybersecurity and privacy online and often take measures to protect themselves from the misleading or malicious consequences of a deepfake. The respondents noted that they do not feel helpless when coming across a deepfake and tend to compare online content with other sources to confirm the whether the content is real or fake. The use of reverse image searches is not a common method of detecting deepfakes among the Greeks with users acknowledging that their own actions are not enough against malicious deepfakes. The general notion is that Greece should have a national deepfakes strategy shaped to limit the spread and protect users.

It is crucial to note that the sample may be considered as not representative of the population of Greece. For instance, no metrics of age were considered in the questionnaire. In addition, only English-speaking Greeks were challenged with deepfakes detection and asked about their opinion and experiences on social media and Internet. The results of the questionnaire show that the participation of males was significantly lower than the females. Specifically, only 20 out of 123 responders declared themselves as male. This translates into 16.3%. Responses for “Female” were 101 (82.1%). Those details can also be viewed in the following Figure 17.

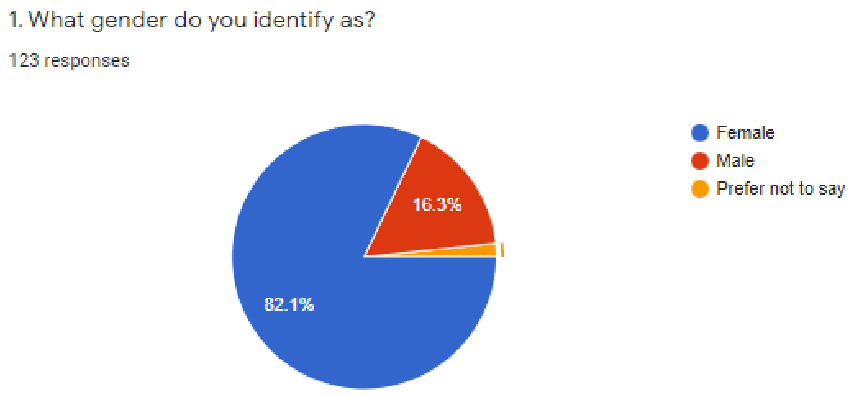


Figure 17. Gender distribution

Currently there is not much confidence in the Greek government's efforts to combat the problem of malicious deepfakes online. This was confirmed with the analysis on the legislation and actions on deepfakes, as described in the Annex A. However, Greeks recognise that the government should not be the only agency responsible for the fight against malicious deepfakes. This issue requires a joint effort with the assistance of the EU, social media companies and educational institutions. Social media companies, such as Facebook, rely on technology-based software to prevent the spread of deepfakes on their platforms.

RQ4 addressed the effectiveness of detection methods used in Greece. Specifically, two different approaches were examined. These were the FB-MSU tool based on image reverse search and AI, and the Ellinika Hoaxes fact-checking and content-moderating activities.

The results presented by Facebook (70% of accuracy) and from the test conducted on the Ellinika Hoaxes team members (81.3% accuracy) indicate that those methods are quite effective. It is also known that Facebook is recently being supported by Ellinika Hoaxes to filter and ban deepfakes on the platform. Thus, a mixed and complimentary approach to fight deepfakes is opted for as is provides the most comprehensive and effective protection.

However, most of those interviewed do not feel confident towards Facebook's deepfake strategy. In general, most users believe that technology-based methods are not the only reliable and effective tool to combat deepfakes. They express the belief that user education and computer literacy can raise cybersecurity and compliment the effort of combating the spread of deepfakes. Organisations such as Ellinika Hoaxes, an online community where users exchange knowledge and collaborate to spot fake news and contribute to the fight against deepfakes, seem to be widely unknown to Greeks. Finally, the legislation is another element integral to the fight against malicious deepfakes. Based on the results obtained from the questionnaire, Greek society feels that not all deepfakes should be banned. Laws should regulate the use of deepfakes to prohibit their use for illicit purposes.

**RQ4** was "What is the effectiveness of deepfakes detection methods used in Greece and their perceived contribution to the cybersecurity of the Greek online community?". **The analysis performed within this thesis shows that those tools are effective, yet they are not considered as such by Greek online community**.

In summary, mitigation of the deepfakes threat is a complex, multifaceted and challenging task faced by Greece. The current threat landscape has shown that deepfakes are not malicious and are often used in the film and entertainment industries, as well as in celebrity gossips and news. However, societal worries and insecurity grow. While users possess some knowledge on detecting deepfakes, they are unable to correctly detect all deepfakes. Cybersecurity threats and attacks, such as online frauds and phishing with weaponised deepfakes have been already experienced by some users from Greece. In the context of deepfakes, actions taken within Greek social media and Internet are not considered sufficient to cope with the escalating threat posed by deepfakes. It is accepted that the task to combat deepfakes requires actions from the Greek government and legislative bodies. It also surpasses the national level and necessitates the participation of the EU, social media companies and educational institutions. The final **RQ5** concerns the issue of a Deepfakes strategy of Greece. Suggestions are made in the following Section 6.1.

## 6.1 Greek National Deepfakes Strategy

The questionnaire results combined with the analysis and discussion surrounding the replies are a stimulus for a discourse on a potential Deepfakes Strategy of Greece. The representatives of the Greek online community are not in favour of the Greek government being the sole organisation responsible for halting the spread of malicious deepfakes. Cooperation with the main social media companies should be part of the national strategy for deepfakes. This way, the Greek government would have access to data on the deepfakes threat in Greek social media, the latest technologies, and strategies to halt the spread of malicious deepfakes. Based on the users' experiences, the platforms that should be of priority are Facebook, Instagram, and YouTube. Greek users indicated that those platforms contained the most images, which are considered by them to be deepfakes. Those platforms are also the most popular within Greek society. Filtering of content, user awareness campaigns, information exchange and cross-platforms cooperation are possible actions to be undertaken.

A crucial aspect in terms of fight against deepfakes is the definition of their appropriate use cases. Deepfakes are used in several contexts and for different purposes. Based on the findings of the questionnaire, Greek society believes that most deepfakes are not malicious. At the same time, there were only a few instances of malicious deepfakes used in cyber attacks such as phishing and frauds.

Thus, the acceptable use of AI and deepfakes should be delineated in legislation. This would help users understand which deepfakes are considered malicious and highlight their negative impact. At the same time, this helps social media platforms and websites to outline their deepfakes strategy for reporting and blocking content.

The lead Government department responsible for mitigating the threat of deepfakes could be the Ministry of Digital Governance. This body is slowly becoming more active in the fields of digital transformation, cybersecurity, Artificial Intelligence, machine learning and virtual reality [131]. Countering malicious deepfakes could be another area of its responsibility. At the same time, non-governmental bodies that are active in cybersecurity issues in Greece should also be involved in raising user awareness on of the rising threat of deepfakes. Examples of such entities are Cyber Alert and Cyber Security International Institute (CSII) [132] [133]. Cyber Alert educates users on cybercrimes and enables reporting them [132]. The categories that are monitored are scams, phishing, and frauds. The entity could expand the education and monitoring to malicious deepfakes. CSII promotes education and awareness and enhances practical activation of citizens in matters of new technologies, internet security, and safe use of social media and cybercrimes.  Finally, it promotes scientific research on cyber threats

[133]. One of the topics which should be added to the curriculum of this organisation is deepfakes and Artificial Intelligence.

User awareness is inseparable to the increase of public's resilience to misinformation and manipulation caused by malicious deepfakes. It also requires the safeguarding of digital literacy and safe Internet and social media use. Besides Greek government and non-governmental organisations, the EU should also be involved in the fight against malicious deepfakes in Greece. This desire was expressed by most participants of the survey. EU law and actions against deepfakes are still being shaped. The activity of the EU and Greece should be complimentary to each other and collectively address any deepfake issue arises.

The use of deepfakes is still rarely used as an attack vector in Greece. Thus, cybersecurity teams and experts have the advantage of being able to prepare defences and tools to mitigate this potential future threat. Greece should also invest in researching forensic techniques for detecting deepfakes, which should be incorporated into the national strategy. This strategy should envision partnerships with schools, universities, research centres, libraries. Also, companies, which deal with Artificial Intelligence should be included as well. Such collaboration would bring together all disciplines of cybersecurity, law, information technologies, education, and artificial intelligence. An official strategy should seek to shape a culture where people pause for a while before sharing content online and spread misinformation.

# 7. Areas for future research

The level of cybersecurity within Greek organisations is not high despite users' digital literacy and cybersecurity awareness increasing [134]. As of writing, Greece has yet to reach a comparative level of cybersecurity competence when compared to other EU states [135]. Estonia is an example of a highly digitalised economy and society. It is often characterised as an international cybersecurity leader [136]. Future research could focus on comparison of those two states in relation to detection of deepfakes by the two societies and how digital literacy and skills impact it. It would also investigate the actions taken within the states to safeguard Internet and social media from malicious deepfakes.

Another proposal for further research is the demographics and their correlation to deepfakes detection ability and experiences on social media and Internet. Age, gender, occupation, education, computer skills are factors which potentially impact the skill to observe deepfakes indicators. Social media and the Internet engagement of users can also be dependent on the aforementioned factors. These metrics were not collected within the scope of the present thesis due to time constraints.

Additionally, a significant point to address in future work is the impact of language on deepfakes detection. The questionnaire included in the thesis was distributed to Greeks who speak English. Future research could investigate the detection skills, opinions, and experiences of non-English speaking Greeks. Such study could provide artifacts on how the language influences the understanding of deepfakes, Internet and social media content.

Finally, future research would incorporate a comparison of sophisticated deepfakes and cheap fakes. The results of the questionnaire showed that sophisticated deepfakes are not more deceiving that cheap fakes. More tests are required though to show if sophisticated deepfakes are trickier and more difficult for humans to detect.

# 8. Conclusions

The findings of this thesis serve to highlight the growing, yet underestimated, threat of deepfakes in Greece. The Greek government has yet to prepare tools, policies, law, and other measures to prevent the spread of malicious deepfakes online. The most significant measures to combat deepfakes have been taken by Facebook as well as independent bodies with no governmental affiliation such as the content-checking organisation Ellinika Hoaxes. Greek society has already experienced illicit deepfakes, though not to a great extent. The current landscape is prevailed by deepfakes related to gossip, celebrity news and entertainment.

Creation of a believable deepfake is within the abilities of any person who knows how to exploit Internet resources. Multiple tools and instructions to generate deepfake content are readily accessible online and easily understandable. However, the creation of highly sophisticated deepfakes demands more experience and processor power and graphic capabilities sufficient to handle the demands of creating deepfakes. The accessibility of the sources to create deepfakes can be concluded to be a potential threat for increasing the number and reach of malicious deepfakes on Greek social media and Internet.

The detection of deepfakes by humans is a challenging task. As the results of the questionnaire showed, highly sophisticated deepfakes are not detectable by most humans. Only 39,8% of the questionnaire responders detected correctly a highly sophisticated deepfake. At the same time, often reviewing manipulated content and being aware of the typical signs of such content as well as tools to detect can significantly empower the individual to combat malicious deepfakes online.

The threat of deepfakes is a relatively new topic. Its different aspects and implications have not been addressed by Greece and the EU. The current state of law both in Greece and the EU are not yet mature enough to handle the matter of malicious deepfakes. There are some legislative tools which could find application in deepfake cases, but a specific piece of law that addresses directly deepfakes is not yet been introduced. In relation to law, the topic of deepfakes requires a reassessment of the current arrangements. This requires a different approach and perspectives on the regulations for copyrights, misinformation, Artificial Intelligence, cyber crimes and cyber security.

Thus, deepfakes are a complex and challenging topic for legal regulation, as it seeps through many fields of legal interest.

It should be noted that there are continuous technological developments in the private sector to develop detection tools with the highest accuracy. There are numerous different approaches to the deepfakes detection. Facebook combines two strategies that aim to constrain the spread of malicious deepfakes online: human detection and image reverse search and AI-based algorithms. This could indicate that such compound can reach the highest possible moderation of malicious manipulated and AI-generated content. This strategy could drive the formulation of a national deepfakes strategy. A combination of legal tools, societal awareness and computer literacy and detection software can provide a comprehensive defense against the threat of deepfakes.

This thesis has aimed to provide a detailed overview of the deepfakes landscape in Greece. The process included a review of the current legislation enforced in Greece and an examination of the societal perceptions and experiences with deepfakes on Greek social media and Internet. It also tested users' ability to detect deepfakes. Finally, it incorporated a discussion on two deepfakes detection methods used in Greece and a suggestion for a national strategy to contain the spread of malicious deepfakes online. The growing availability of deepfakes generative tools will come coupled with risks. This includes threats to social order, economy, democratic institutions, and cybersecurity. It is vital that an effective coalition between government and the commercial sector to be shaped and develop appropriate regulations and safeguards before more incidents with the use of deepfakes occur.

Deepfakes can be considered as a growing problem within the Greek social media and the Internet and as such is a new cybersecurity threat for Greek users. Despite this, the spread so far is limited, although there have already been cases of reported. Both users and Ellinika Hoaxes have lodged cases of malicious deepfakes leveraged for misinformation and cyber crimes. This is a concern for Greek government, as the occurrence of such events is expected to increase in the near future. It would be advantageous for Greece to shape a strategy to counter the threat of deepfakes that could prevent the spread of malicious deepfakes. This would regulate their use and enhance Greek users' defences against this increasing cybersecurity threat.

# Annexes

## Annex A. Deepfakes Legislation

### Greek legislation on deepfakes

Legislation specific to deepfakes does not exist as such in Greece. There is also no law that regulates the use of Artificial Intelligence in the creation of material. This creates an opportunity for other legal frameworks and doctrines to fit in for the regulation of deepfakes. Legal bodies and experts often mention that copyright laws could address the issue of malicious deepfakes [137] [138]. The creation of a deepfake usually involves the manipulation of existing audio-visual content, which might be protected by copyright law. In this case, the permission of the copyright owner must be obtained to use this content as source for a deepfake.

In the case of Greece, this means that the existing Copyright Law 2121/1993 should be investigated. According to the Article 2 of the law, which is written in English, the objects considered as intellectual creation and thus protected are:

***"1.*** *The term work shall designate any original intellectual literary, artistic or scientific creation, expressed in any form, notably written or oral texts, musical compositions with or without words, theatrical works accompanied or unaccompanied by music, choreographies and pantomimes, audio-visual works, works of fine art, including drawings, works of painting and sculpture, engravings and lithographs, works of architecture and photographs, works of applied art, illustrations, maps and three-dimensional works relative to geography, topography, architecture or science.*

**2.** *The term work shall, in addition, designate translations, adaptations, arrangements and other alterations of works or of expressions of folklore, as well as collections of works or collections of expressions of folklore or of simple facts and data, such as encyclopaedias and anthologies, provided the selection or the arrangement of their contents is original. Protection afforded to the works listed in this paragraph shall in no way prejudice rights in the pre existing works, which were used as the object of the alterations or the collections.*

***2a.*** *Databases which, by reason of the selection or arrangement of their contents, constitute the author's intellectual creation, shall be protected as such by copyright. The copyright protection shall not extend to the contents of databases and shall be without prejudice any rights subsisting in those contents themselves. Database is a collection of independent works, data or other, materials arranged in a systematic or methodical way and individually accessible by electronic or other means.*

***3.*** *Without prejudice to the provisions of Section VII of this Law, computer programs and their preparatory design material shall be deemed to be literary works within the meaning of the provisions on copyright protection. Protection in accordance with this Law shall apply to the expression in any form of a computer program. Ideas and principles which underlie any element of a computer program, including those which underlie its interfaces, are not protected under this Law. A computer program shall be protected if it is original in the sense that it is the author's personal intellectual creation.*

***4.*** *The protection afforded under this Law shall apply regardless of the value of the work and its destination and regardless of the fact that the work is possibly protected under other provisions.*

***5.*** *The protection afforded under this Law shall not apply to official texts expressive of the authority of the State, notably to legislative, administrative or judicial texts, nor shall it apply to expressions of folklore, news information or simple facts and data."* [139]

Based on the above, the current Greek law does not explicitly refer to deepfakes generally or to products generated or manipulated with Artificial Intelligence. It is also valuable to note that the creation of deepfakes for education purposes or scientific content requires permission of the copyright owner of the source audio-visual content.

Greece has expressed a strengthening interest in the quick transposition of the Article 17 of the Directive 2019/790 on copyright and related rights into national law. A drafting committee that would enable the smooth integration of the directive into the Greek legal system was also created [140]. This Directive aims to regulate the copyright issues in the newly shaped and unknown environment of the Digital Single Market and information society [141]. The directive is a response to the ongoing changes in the social media and Internet domain. The way that copyright content is now being shared, accessed, downloaded requires new instruments and answers from legal bodies [141].

Deepfakes are a product of AI- based technologies. Thus, rules and regulations for the use of Artificial Intelligence could have significant application on malicious deepfakes. Greece is a state which does not regulate the use of Artificial Intelligence. In 2021, the Minister of State and Digital Governance, Pierrakakis announced that Greece is ready to present a National Scheme for Artificial Intelligence. The aim is to prepare Greece for the 4th technological revolution, but also ensure that Greek society benefits from the use of Artificial Intelligence [142] . He also highlighted that for now as Greece lacks its own legislation, the country tries to follow and incorporate the rules and laws passed on the EU level into the national law.

In late September 2021, the Ministry of Justice passed a decision to replace law 191 of the Penal Code regarding fake news. The amended legislation states that that:

*"Anyone who publicly or via the internet spreads or disseminates in any way false news that is capable of causing concern or fear to the public or shattering public confidence in the national economy, the country's defence capacity or public health shall be punished by imprisonment for at least three months, and a fine.*

*"If the act was repeatedly committed through the press or via the internet, the perpetrator is punished with imprisonment of at least six months and a fine. The actual owner or issuer of the instrument with which the acts of the previous paragraphs were performed is punished with the same penalty.*

*"Anyone who through negligence is guilty of any of the acts of the previous paragraph shall be punished by imprisonment of up to one year or a fine."* [143]

The above piece of legislation could apply in cases of deepfakes. The challenge which the courts face is to prove that the actor behind the generation and initial spread of a malicious deepfake aims to misinform the public opinion and foment civil distress and unrest and undermine the order. At the same time, this law is vague enough to create some concerns over free speech and expression of opinion. According to many reporters, this could hinder journalism in Greece [144].

Another relevant piece of law that could assist the Greek courts in addressing the malicious use of deepfakes is the criminal code. The main legal instrument is the Greek Criminal Code, *1805/1988,* which does not refer exclusively to cyber crimes committed with the use of deepfakes. However, the laws regulating some categories of cyber crimes in Greece are general and vague in nature. This creates a potential for interpretation and discussion. Deepfakes are a means to commit financial frauds, phishing, and identity theft in the cyber domain.

According to the Article 386A of the Greek Criminal Law,

*"Whoever, with the purpose of gaining illegal profit, damages foreign property by influencing by any means of data processing, faces a penalty up to 10 years' imprisonment and a penalty fee. Apart from the above-mentioned case, identity theft can constitute several criminal offenses under the Greek Criminal Code, depending on the manner and reason for which the offender obtains access to identity data."* [145]

The Article 386A paragraph 1 of the Greek Criminal Code, foresees regulations for phishing attacks. Specifically, it stipulates that:

*"When phishing has the meaning of attempting to fraudulently acquire through deception sensitive personal information (such as passwords), it falls under Art. 386, par. 1 of the GCC and bears a penalty of 10 days to five years imprisonment and a penalty fee.*

*On the contrary, if the phishing is defined as a type of fraud that involves the use of a computer, by creating false digital resources to resemble those of legitimate entities, to induce individuals to reveal or disclose sensitive personal information, then it falls under Art. 386A, par. 1 of the GCC and bears a penalty of 10 days to five years imprisonment and a penalty fee.*

*In both cases, when the damage that occurred as a result of phishing exceeds the amount of €120,000, the penalty is imprisonment of up to 10 years and a penalty fee."* [145]

The afore-mentioned Article 386 addresses also the crime of "Fraud with a computer". In details:

*"Whoever, with the intent of obtaining for himself or for a third person an unlawful material benefit, damages the assets of another, by affecting the elements of a computer either through incorrect configuration of a program or interference in the operation of a program or use of incorrect or incomplete data or in any other way, shall be punished with the punishments of article 386. The damage of the assets exists even if the victims cannot be located. For estimation of the amount of the damage it is irrelevant whether the victims are one or more persons."* [146]

In the light of the above, the social media and Internet environment are simply the platforms for conducting cyber crimes and deepfakes are a weapon used by malicious actors. While the act of the creation of a deepfake cannot be punished, the use of such content for illicit purposes and computer crimes could be punished. The existing legislation is not referring directly to deepfakes. In the case a deepfake is used within the context of committing a cyber crime, punishment for the perpetrator is possible. The act of creating malicious deepfakes is not a considered as a crime and thus not punishable in Greece.

Another important point is that the legislation of Greece does not deal with the deepfake pornography. Pornography, except for child pornography is legal [147]. Thus, Greece has no legal instruments which address the problem of malicious deepfakes pornography and deals with the negative impact on the victims.

Even if some of the above laws could be applied to deepfake cases, there are some elements to consider by courts. It is often difficult to detect the perpetrator of a cyber crime with the use of a deepfake. Malicious actors often act anonymously, use technology to hide their traces and identity. They cannot be easily identified and held accountable for their actions online. Cyber criminals can act from a foreign state and thus the legal frameworks and law enforcement of Greece does not have the jurisdiction and authority to act upon cyber crime.

In summary, the current legal ecosystem in Greece does not directly address the matter of deepfakes or Artificial Intelligence as legal challenges. Some laws could be interpreted in such a way that could encompass crimes with the use of deepfakes. However, no law deals with the aftermath of a malicious deepfakes in the cyber domain. For instance, manipulation and misinformation of society, erosion of public trust towards media, financial impact due to extortion, fraud, or identity theft. The legal route for the victims remains challenging.

In recent months, there have been some encouraging efforts from the Greek government to deal with the spread of malicious fake content online and the Artificial Intelligence. The legal landscape is progressing to incorporate novel technologies and re-evaluate the application of existing laws and principles. The progress in Greece is visible. The above plans, decisions and structures formed are small steps towards creating a legal framework for AI and its subcategories such as deepfakes. While Greece formulates its initial strategy, some important legal events took place on the level of the European Union. Those developments concern Greece as well, as the country is an EU member state.

## Deepfakes legislation of the European Union

Greece is a member of the European Union. Due to that, laws and decisions taken by EU legislative bodies also have implementation in Greece. The current European regulatory landscape does currently and directly address the matter of deepfakes. However, EU has made a significant step to regulate AI. Specifically, in April 2021, the European Commission published the "Proposal for a regulation laying down harmonised rules on Artificial Intelligence" [148]. This law aims to allow use of deepfakes. It lays down some conditions for it. The first one is the obligation for transparency from the deepfake creator. According to the Article 52(3):

*"Users of an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake'), shall disclose that the content has been artificially generated or manipulated."* [148]

The exception to this is:

*"However, the first subparagraph shall not apply where the use is authorised by law to detect, prevent, investigate and prosecute criminal offences or it is necessary for the exercise of the right to freedom of expression and the right to freedom of the arts and*

*sciences guaranteed in the Charter of Fundamental Rights of the EU, and subject to appropriate safeguards for the rights and freedoms of third parties."* [148]

Another important element is that the regulation aims to ban only the use of AI which is considered as "high risk" and that "poses an unacceptable risk to the safety and fundamental rights of EU citizens" [148].

Based on the above, it becomes apparent that the EU takes into consideration the beneficial use of deepfakes and AI in general. Thus, science, education, and "art" deepfakes are considered as beneficial and are exempted from any obligations for the creator. The promotion of technology and advancement in the EU is not halted. The aim is to provide legal clarity on the use and development of AI. This way, the EU can ensure that AI is developed and used in respect to its fundamental rights and safety requirements and offer lawful, safe, and trustworthy AI applications.

The weakness of the AI proposal is that it does not stipulate the measures and form of punishment for those who do not comply with the transparency obligations. Also, the categories of exceptions are too broadly formulated, which could enable malicious deepfakes to be exempted. There is a need for more clarifications and detailed examples on certain aspects of this proposal.

Another piece of legislation relevant to deepfakes is the General Data Protection Regulation (GDPR). According to Article 4 (1):

*"'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;"* [149]

This means that if a deepfake depicts a natural person, or if the process of deepfake creation uses elements related to a natural person, this can be considered as personal data. Based on the above, this law is applicable from the moment of deepfake creation with the use of software and training material to the moment of the deepfake use and/or dissemination.

According to Chapter 2, Article 6 (1), there are six legal grounds which characterise personal data processing as lawful, fair, and transparent. Specifically:

a) Consent,
b) Performance of contract,
c) Legal requirement,
d) Vital interest,
e) Public interest,
f) Legitimate interest [150].

In the context of deepfakes creation and use, the points a and f are the most relevant. According to the Chapter 2, Article 7, consent must be provided by the data subject permitting his/her personal data processing. The consent must be clear, given freely and easy to withdraw [151]. Legitimate interest is as stated in Article 9 refers to whenever a third party uses personal data in a way that the data subject would expect [152]. The personal data processing is not required by the law and needs to have a clear and specific intended outcome or benefit to the third party [152]. The term "legitimate interest" is broad and could incorporate the use of personal data for the creation and dissemination of deepfake content without this being unlawful. A deepfake used for political commentary and satire by a news outlet is an example of this.

The deepfake is created with the manipulation of source content. This content may be protected by copyright laws. In the EU, there is no harmonisation of the national copyright laws of the member states [153]. In the light of this, the EU does not provide a legal instrument to address deepfakes, while the national copyright law of Greece prevails as the main legal framework for copyright issues.

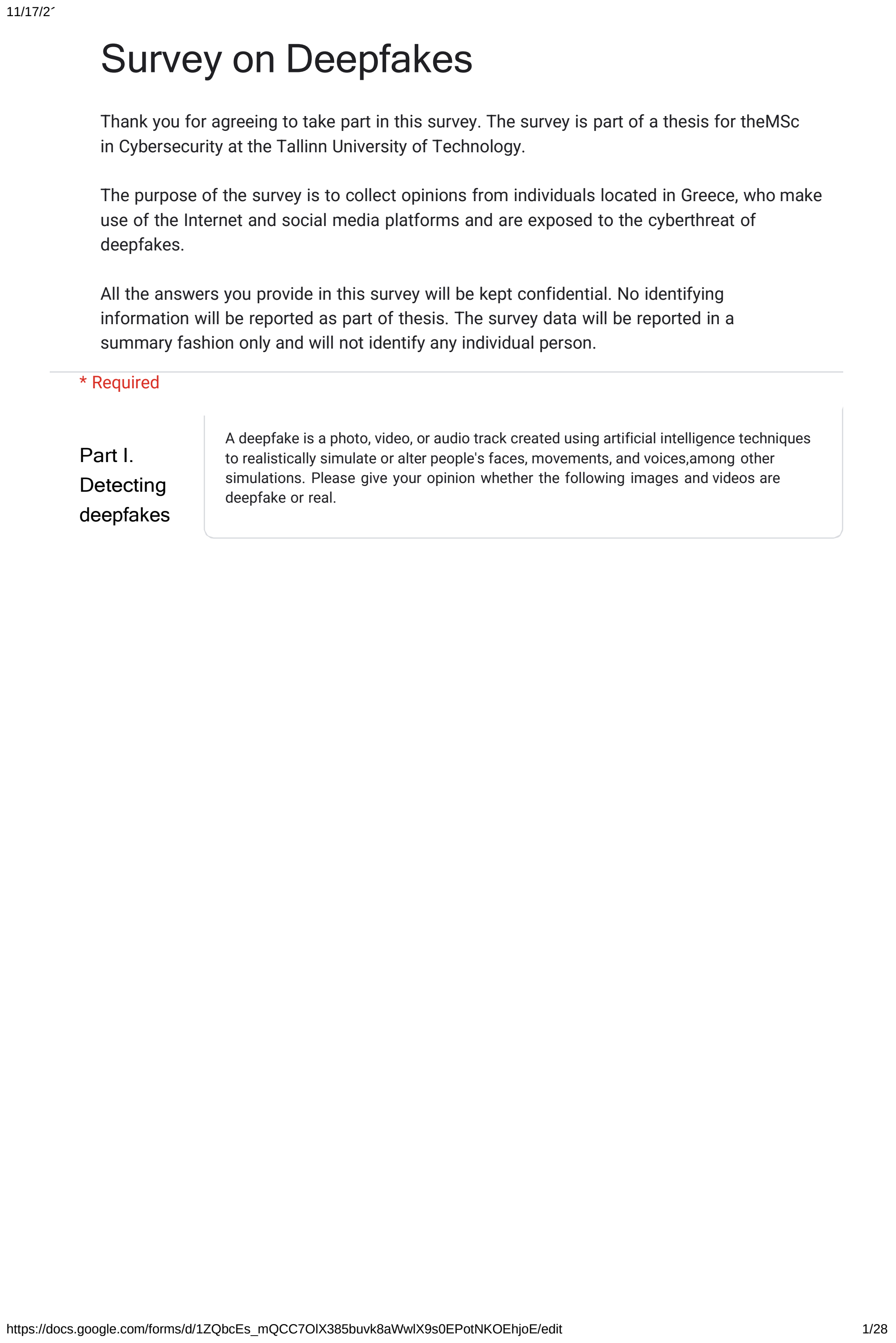
Finally, the Digital Services Act of 2020 aims to "to create a safer digital space in which the fundamental rights of all users of digital services are protected" [154]. The focus is put upon the social networks and content-sharing platforms. The act takes steps on matters of social media speech, illegal online content, disinformation, and copyright issues. All are relevant to the issue of deepfakes regulation. Companies hosting third party's digital content are not liable for it, unless they know that this content is illegal [154]. The act aims to improve the content moderation on social media platforms and create a rulebook of steps to follow when published content is flagged as illegal. At this moment, this act is a proposal by the European Commission, which needs approval by the European Council and European Parliament [155].

Compared to Greek law, the EU law is more detailed and aware of the present problems related to the technological advancements and AI. Greek law could be characterised as outdated. It is not following and responding to the latest technological events which take place in the cyber domain. The EU laws seem more relevant and fitting to the current problems rising from new technologies and digitalisation. It is worth mentioning that in both cases the laws on AI are still in progress of being passed and implemented. However, developments in deepfakes regulation are more advanced within the EU rather than in Greece with the former showing more confidence and decisiveness with its law proposals and plans. In the present time, the EU and Greek law seem to be complimentary to each other in terms of regulating malicious deepfakes. Where one law presents a gap, the other can cover this gap and vice versa.

All in all, this chapter discussed the **RQ1** “What is the current legislation in Greece on deepfakes?”. It was assumed that there is no legislation in Greece directly regulating deepfakes technology. This assumption is correct, as the above analysis proved.

# Annex B. Questionnaire

The questionnaire is available at the following link: https://forms.gle/Bem4K8hjSvPmwSJdA .



11/17/2´

## Survey on Deepfakes

Thank you for agreeing to take part in this survey. The survey is part of a thesis for theMSc in Cybersecurity at the Tallinn University of Technology.

The purpose of the survey is to collect opinions from individuals located in Greece, who make use of the Internet and social media platforms and are exposed to the cyberthreat of deepfakes.

All the answers you provide in this survey will be kept confidential. No identifying information will be reported as part of thesis. The survey data will be reported in a summary fashion only and will not identify any individual person.

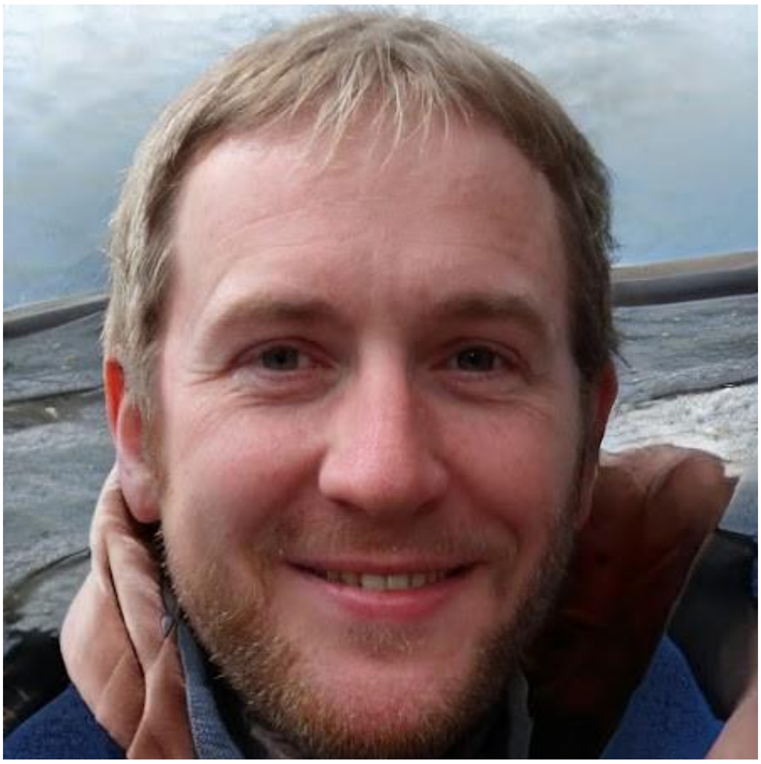
* Required

Part I. Detecting deepfakes

A deepfake is a photo, video, or audio track created using artificial intelligence techniques to realistically simulate or alter people's faces, movements, and voices,among other simulations. Please give your opinion whether the following images and videos are deepfake or real.

https://docs.google.com/forms/d/1ZQbcEs_mQCC7OlX385buvk8aWwlX9s0EPotNKOEhjoE/edit 1/28

1. 1. *

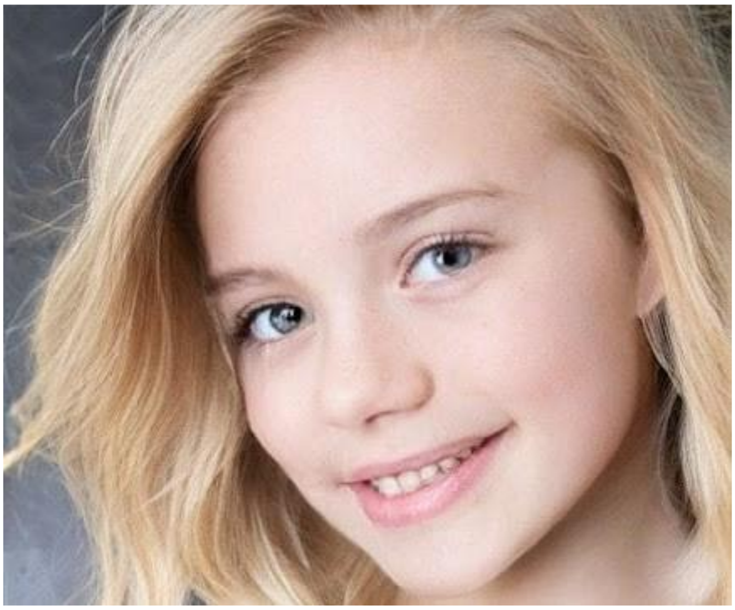


*Mark only one oval.*

[✓ ○ Deepfake]

○ Authentic photo

2. 2. *

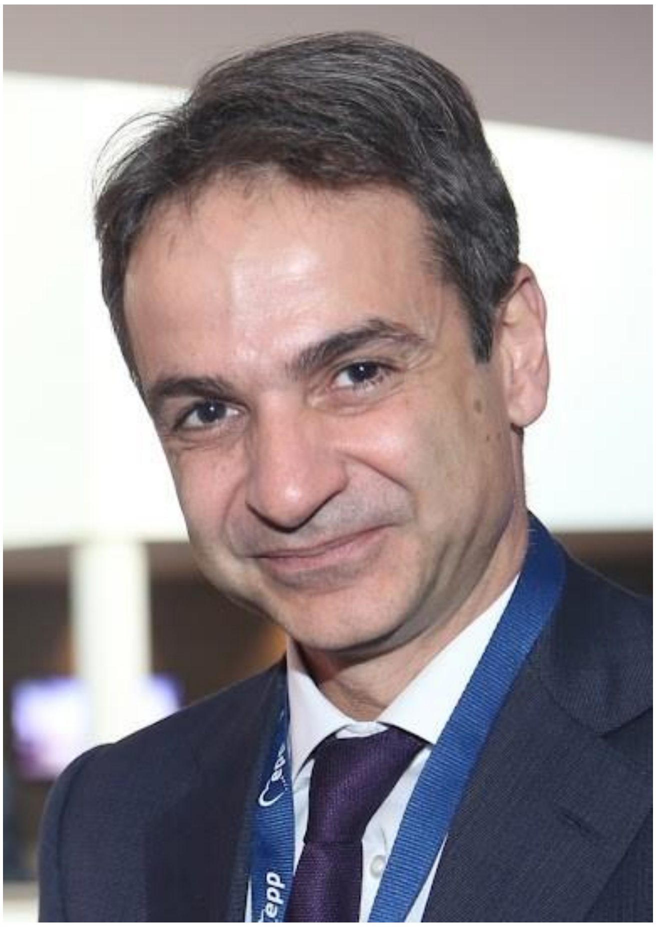


*Mark only one oval.*

○ ✓ Deepfake

○ Authentic photo

3. 3. *

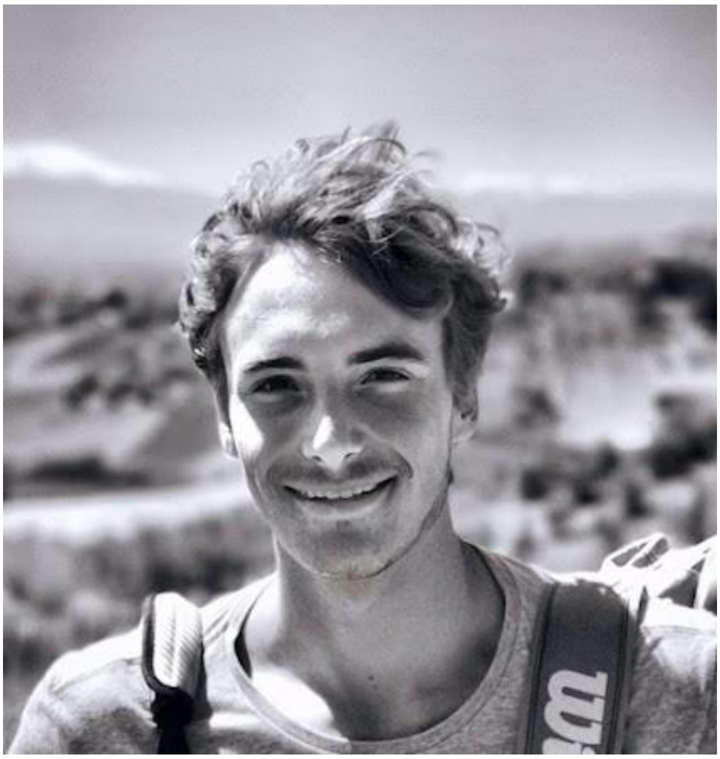


*Mark only one oval.*

○ Deepfake

☑ Authentic photo

4.    4. *

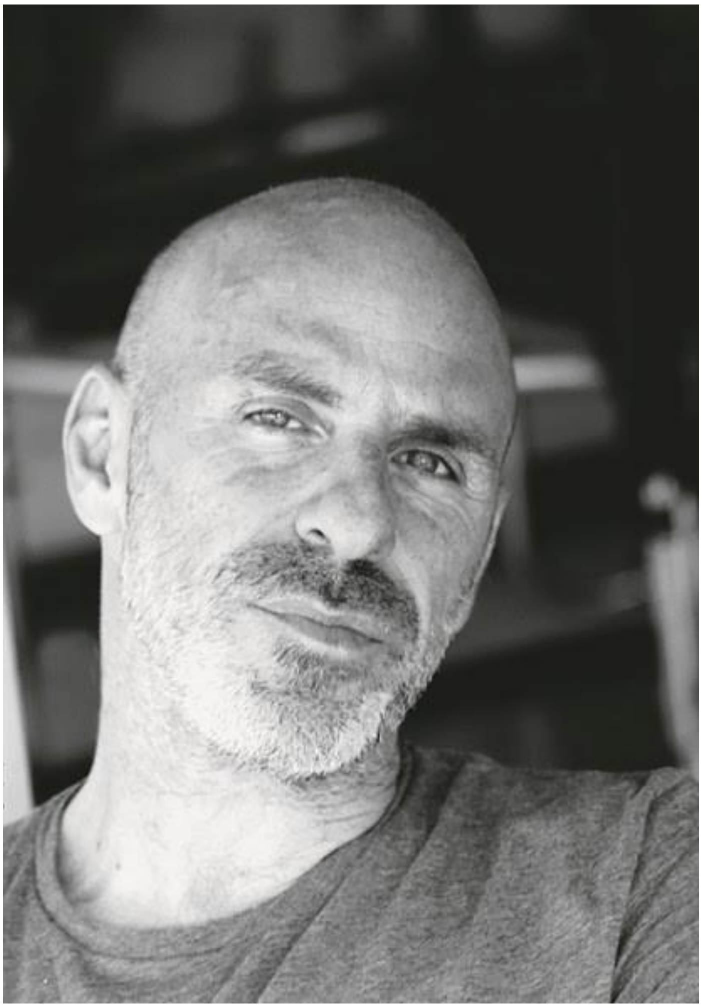


*Mark only one oval.*

☑ Deepfake

○ Authentic photo

5. 5. *

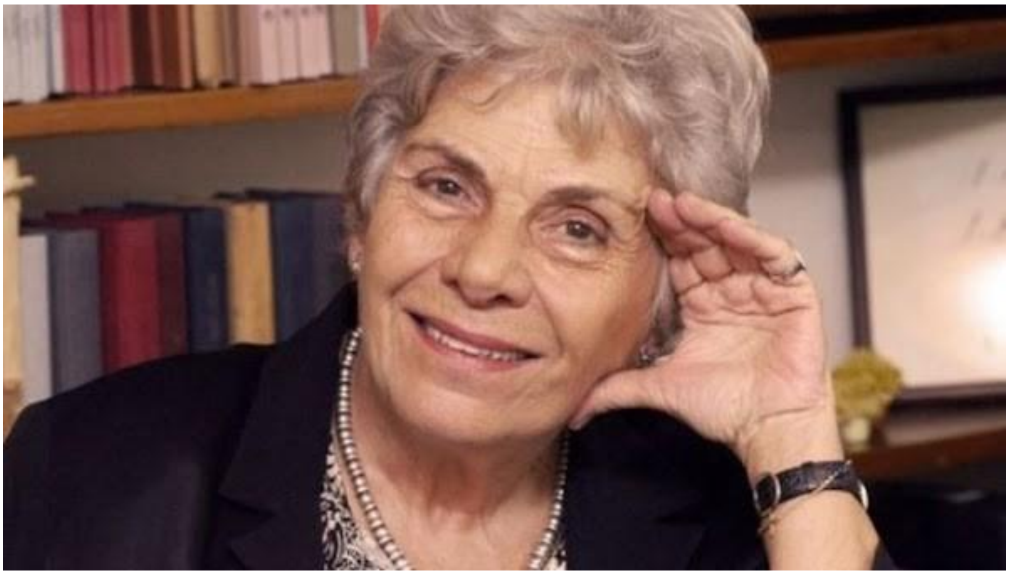


*Mark only one oval.*

○ Deepfake

✓ Authentic photo

6. 6. *

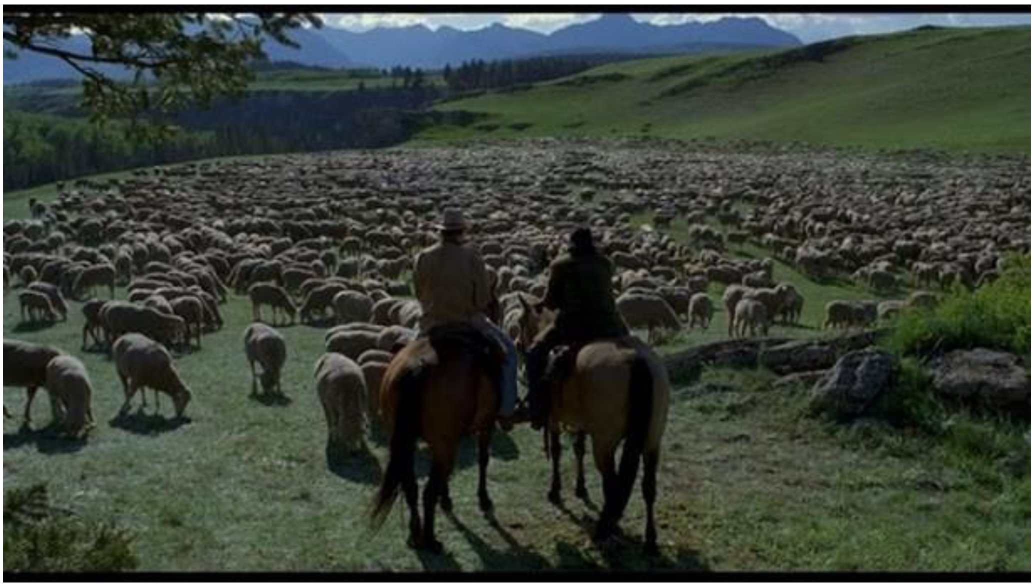


*Mark only one oval.*

○ Deepfake

○ ✓ Authentic photo

7. 7. *

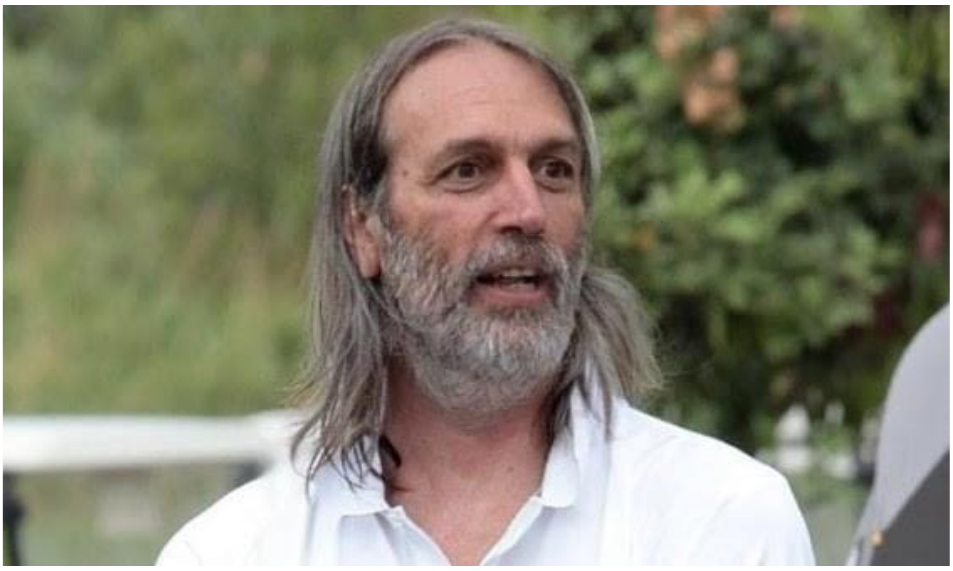


*Mark only one oval.*

○ ✓ Deepfake

○ Authentic photo

8. 8. *

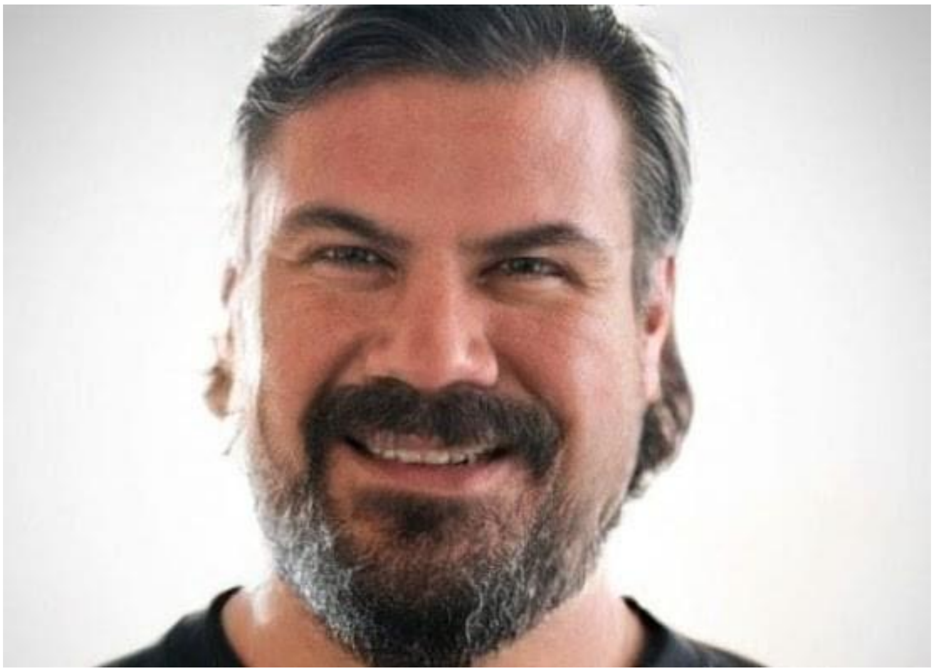


*Mark only one oval.*

☑ Deepfake

○ Authentic photo

9. 9. *

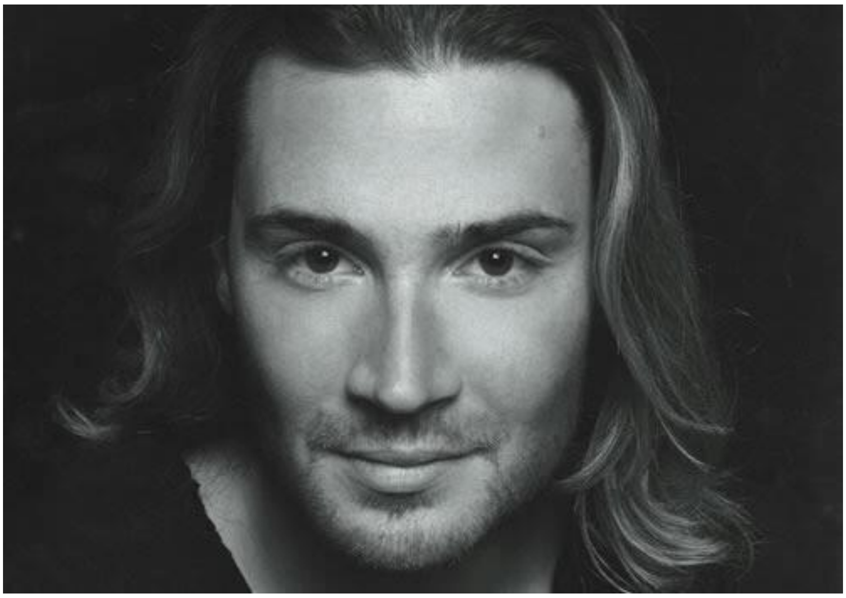


*Mark only one oval.*

☑ Deepfake

○ Authentic photo

10. 10. *



*Mark only one oval.*

○ Deepfake

☑ Authentic photo

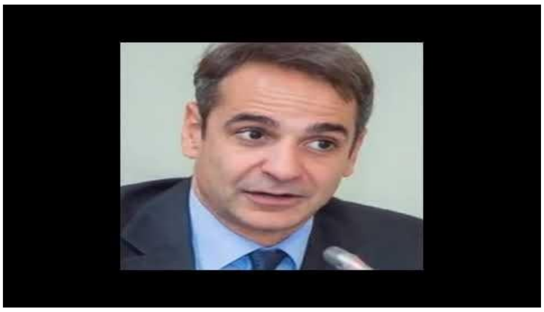
1 of the following videos is authentic.

11. 1. *

*Mark only one oval.*

☑ Deepfake

○ Authentic video



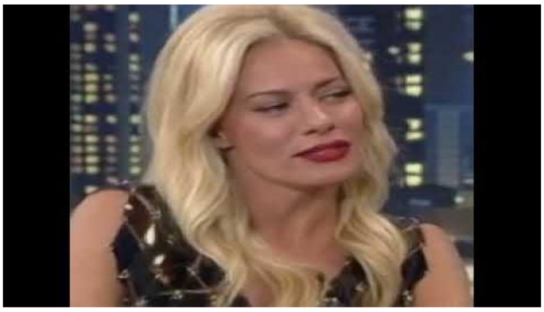
http://youtube.com/watch?v=ynp1zLjoNaU

12. 2. *

*Mark only one oval.*

☑ Deepfake

○ Real video



http://youtube.com/watch?v=32Eb1GtQgpU

13. 3. *

*Mark only one oval.*

○ Deepfake

☑ Real video

http://youtube.com/watch?v=GQWaP8tr8_g

In this section you are asked to provide insight into your experience on social media, the Internet and deepfakes.

Part II.

14. 1. What gender do you identify as? *

*Mark only one oval.*

○ Female

○ Male

○ Prefer not to say

15. 2. How would you rate your computer skills? *

*Mark only one oval.*

○ Basic

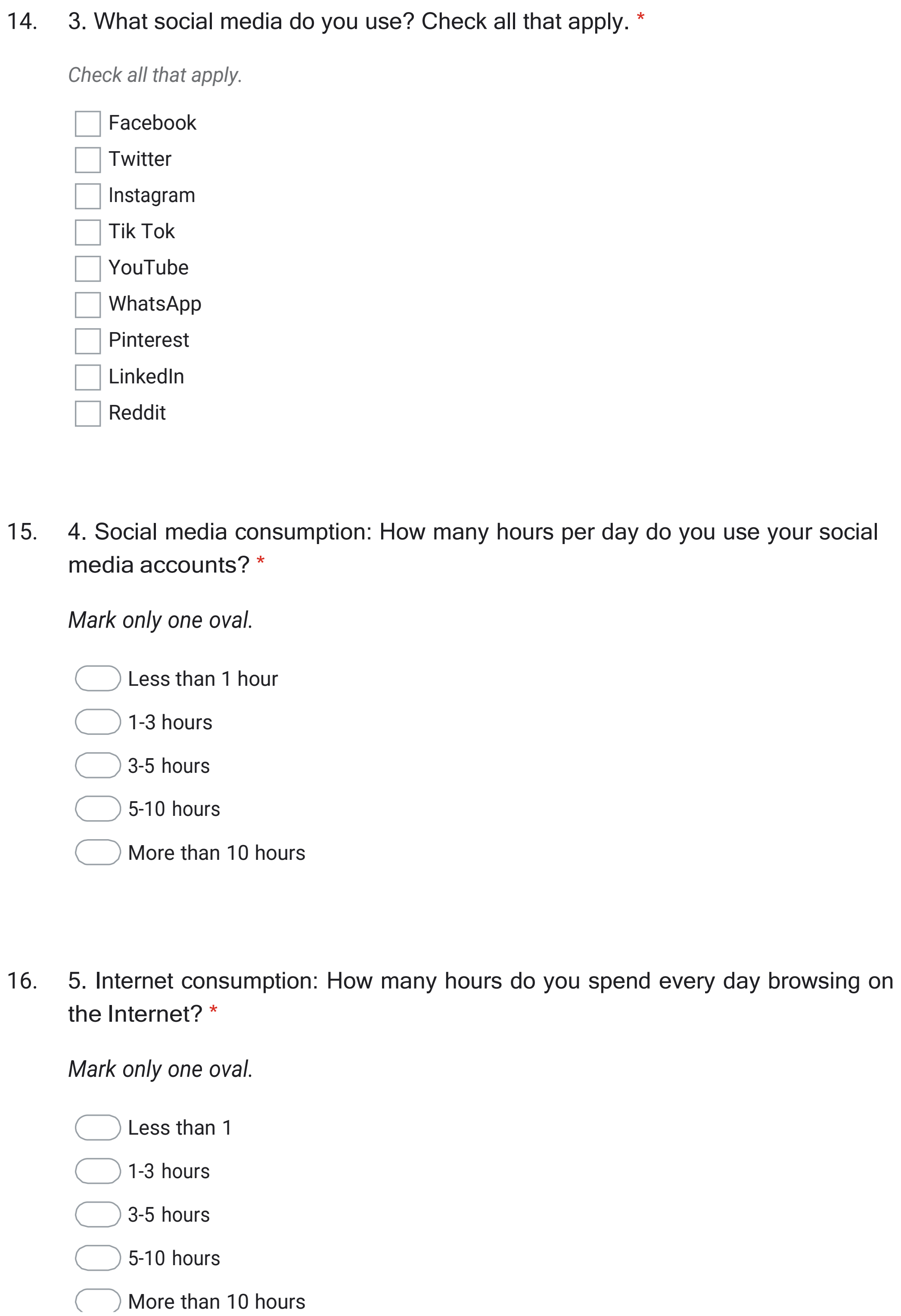
○ Medium

○ Advanced

○ Proficient

○ I do not know

○ None



14. 3. What social media do you use? Check all that apply. *

*Check all that apply.*

□ Facebook
□ Twitter
□ Instagram
□ Tik Tok
□ YouTube
□ WhatsApp
□ Pinterest
□ LinkedIn
□ Reddit

15. 4. Social media consumption: How many hours per day do you use your social media accounts? *

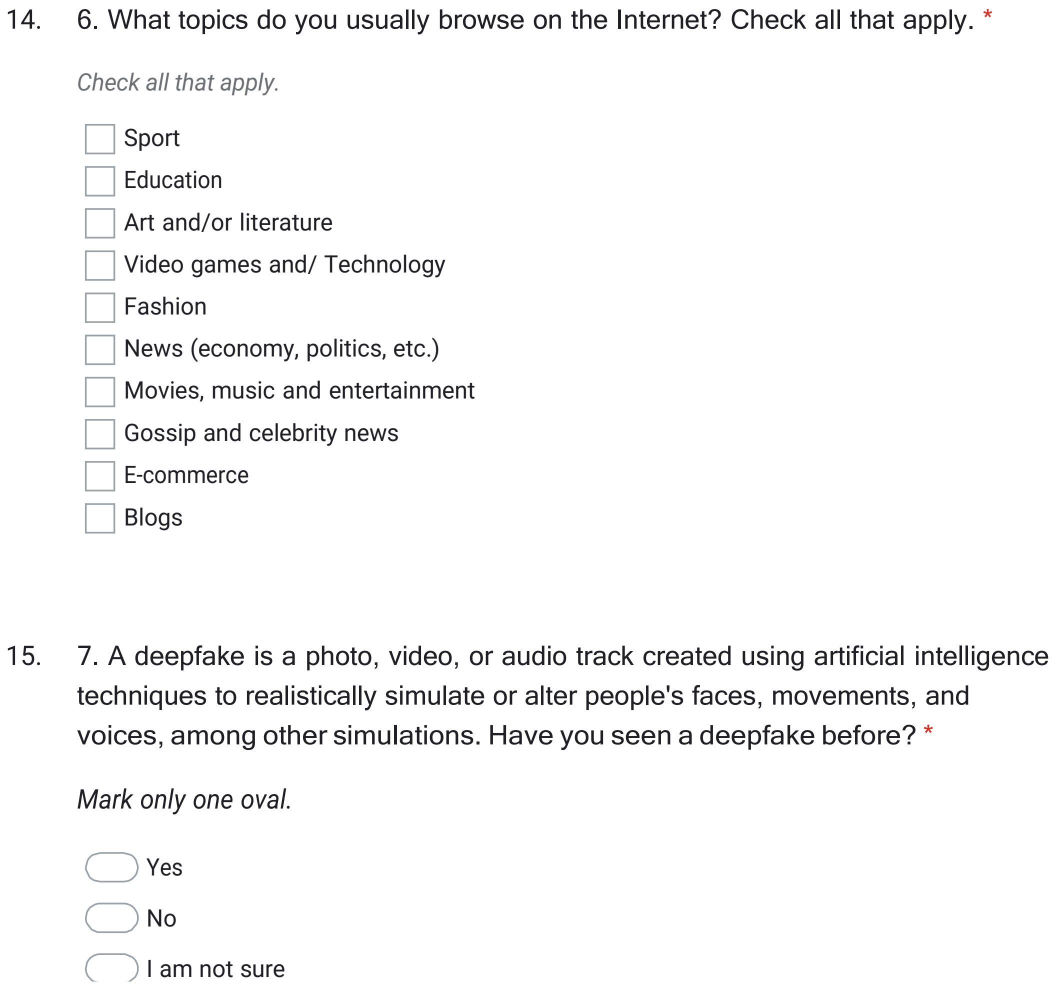
*Mark only one oval.*

○ Less than 1 hour
○ 1-3 hours
○ 3-5 hours
○ 5-10 hours
○ More than 10 hours

16. 5. Internet consumption: How many hours do you spend every day browsing on the Internet? *

*Mark only one oval.*

○ Less than 1
○ 1-3 hours
○ 3-5 hours
○ 5-10 hours
○ More than 10 hours



14. 6. What topics do you usually browse on the Internet? Check all that apply. *

*Check all that apply.*

□ Sport
□ Education
□ Art and/or literature
□ Video games and/ Technology
□ Fashion
□ News (economy, politics, etc.)
□ Movies, music and entertainment
□ Gossip and celebrity news
□ E-commerce
□ Blogs

15. 7. A deepfake is a photo, video, or audio track created using artificial intelligence techniques to realistically simulate or alter people's faces, movements, and voices, among other simulations. Have you seen a deepfake before? *

*Mark only one oval.*

○ Yes
○ No
○ I am not sure

14. 8. If yes, where do you think you have come across deepfakes? Check all that apply. *

*Check all that apply.*

□ Movies
□ Social media filters on apps
□ Art and/or history
□ Voice assistant (e.g., Alexa, Suri)
□ Photoshopped, cropped or manipulated content
□ Memes
□ Dead artists recreated in museums, concerts or movies
□ Satire videos of politicians or celebrities
□ Commercials and marketing materialsVideo
□ games
□ Health
□ I don’t know
□ None of the above

15. 9. How many deepfakes do you encounter daily in social media platforms? *

*Mark only one oval.*

○ None
○ 1
○ 2-5
○ 6-10
○ 10-15
○ More than 15

23. 10. How many deepfakes do you encounter daily while browsing on the Internet? *

*Mark only one oval.*

○ None

○ 1

○ 2-5

○ 5-10

○ 10-15

○ More than 15

24. 11. From the social media you use, in which platforms do you see the most deepfakes? Mark maximum 5. *

*Check all that apply.*

□ Facebook

□ Twitter

□ Instagram

□ TikTok

□ Whatsup

□ YouTube

□ Pinterest

□ Reddit

□ LinkedIn

24. 12. From the websites you visit, in which website category have you seen the most deepfakes? *

*Check all that apply.*

□ Sport
□ Education
□ Art and/or history
□ Video games and/or Technology
□ Fashion
□ News (economy, politics, etc.)
□ Movies, music, entertainment
□ Gossip and celebrity news
□ E-commerce

25. 13. What feelings do you associate with deepfakes? Check all that apply. *

*Check all that apply.*

□ Manipulation
□ Disorientation
□ Joke/Kidding
□ Entertainment
□ Neutral
□ Distrust
□ Worry
□ Danger

24. 14. What is the strongest feeling associated with deepfakes? Check only one. *

*Mark only one oval.*

○ Manipulation

○ Disorientation

○ Joke/Kidding

○ Entertainment

○ Neutral

○ Distrust

○ Worry

○ Danger

25. 17. Which areas of life in Greece do you think are most harmed by deepfakes? Check all that apply: *

*Check all that apply.*

□ Environmental policiesHealth

□ National security

□ Economy and finance

□ Voting decisions

□ Trust in elected representatives

□ Migration policies

□ Trust in public institutions

□ Privacy of individuals

□ Cybersecurity of businesses

□ Cybersecurity of users

Part III.

This questionnaire uses the five-point Likert scale ranging from “strongly disagree” to “strongly agree” 1 Strongly Disagree 2 Disagree 3 Neutral 4 Agree 5 Strongly Agree. Please rate the level of agreement with the following statements:

24. 1. I trust everything posted online and/or social media. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

25. 2. I find online and/or social media content as misleading and harmful for users. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

26. 3. I approach critically the content posted online and/or in social media. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

27. 4. All deepfakes are malicious. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

24. 5. Deepfakes benefit some spheres of life such as education and entertainment. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

25. 6. Most deepfakes are a threat to my cybersecurity. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

26. 7. I feel confident in detecting deepfakes. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

27. 8. I have experienced a phishing attempt with the use of a deepfake. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

24. 9. I have experienced an online fraud with the use of a deepfake. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

25. 10. In relation to deepfakes, I feel secure browsing on the Internet. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

26. 11. In relation to deepfakes, I feel secure while using social media. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

27. 12. In relation to deepfakes, I am worried about my cybersecurity and privacy online. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | | | | | | Strongly agree |

24. 13. When I come across a potential deepfake, I accept that I am helpless, and I cannot trust anything online. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

25. 14. When I come across a potential deepfake, I compare this media with other reports online. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

26. 15. When I come across a potential deepfake, I do a reverse image search to seeif there are any existing photos or videos that might have been used as a basefor a deepfake. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

27. 16. The Greek government is the main responsible for stopping the spread of deepfakes online for my cybersecurity. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

24. 17. I am confident in the government’s efforts to combat the problem of malicious deepfakes online. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

25. 18. Social media platforms are the main responsible for stopping the spread of deepfakes online for my cybersecurity. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

26. 19. Social media companies such as Facebook uses AI-based detection tools to prevent the spread of deepfakes on the platform. I have confidence in the technologies used by Facebook to protect its users from deepfakes. *

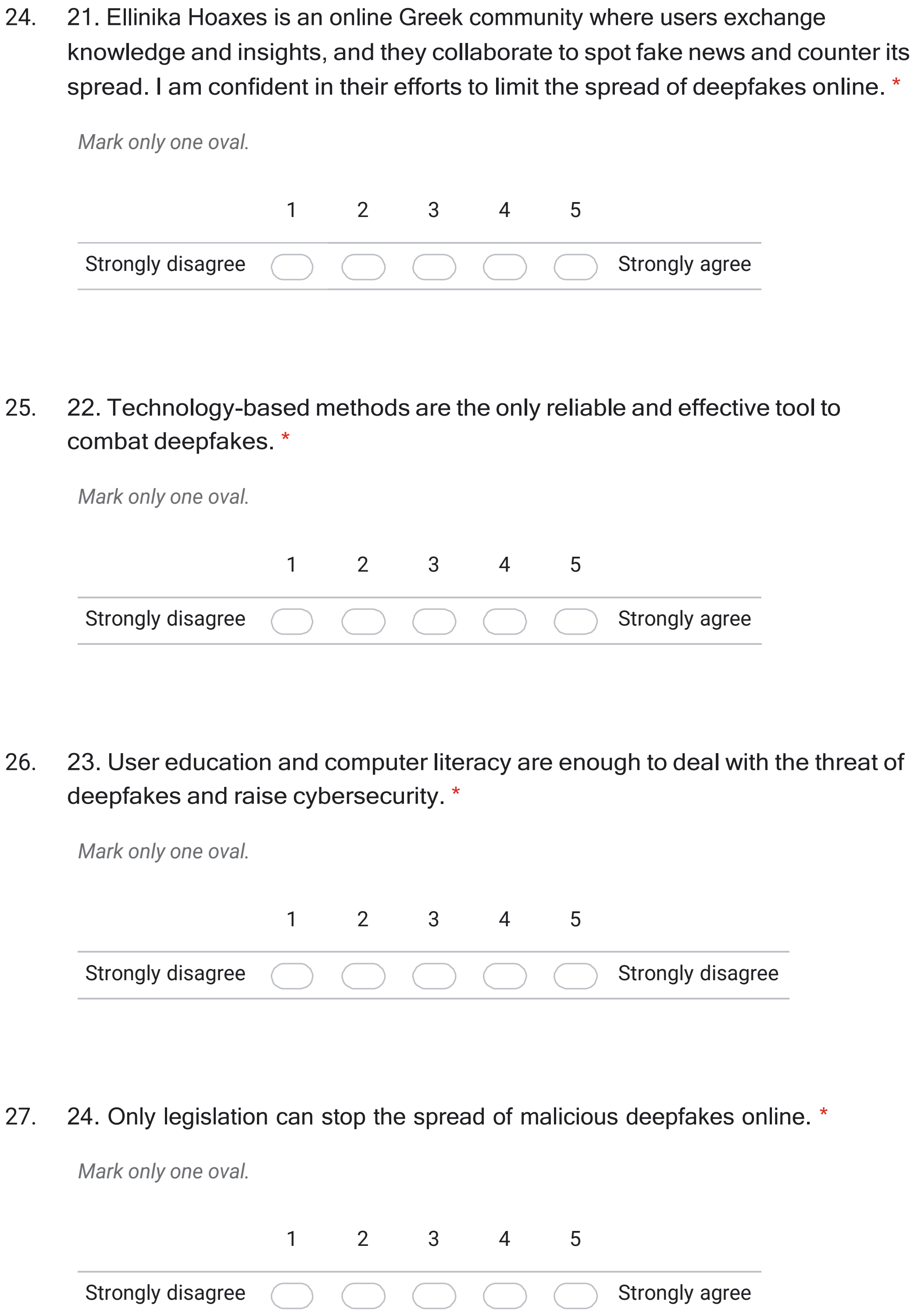
*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

27. 20. Non-governmental and educational organizations are the main responsible to protect users from deepfakes. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |



24. 21. Ellinika Hoaxes is an online Greek community where users exchange knowledge and insights, and they collaborate to spot fake news and counter its spread. I am confident in their efforts to limit the spread of deepfakes online. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

25. 22. Technology-based methods are the only reliable and effective tool to combat deepfakes. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

26. 23. User education and computer literacy are enough to deal with the threat of deepfakes and raise cybersecurity. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly disagree |

27. 24. Only legislation can stop the spread of malicious deepfakes online. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

24. 25. All deepfakes should be banned. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

25. 26. Legislation should not ban all deepfakes. Laws should regulate the use of deepfakes to prohibit their use for illicit purposes. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

26. 27. The European Union should be involved in the efforts of Greece to mitigate the threat of the deepfakes online. *

*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

27. 28. Addressing deepfakes is a responsibility solely of the Greek government. The EU and non-Greek entities should not be involved. *

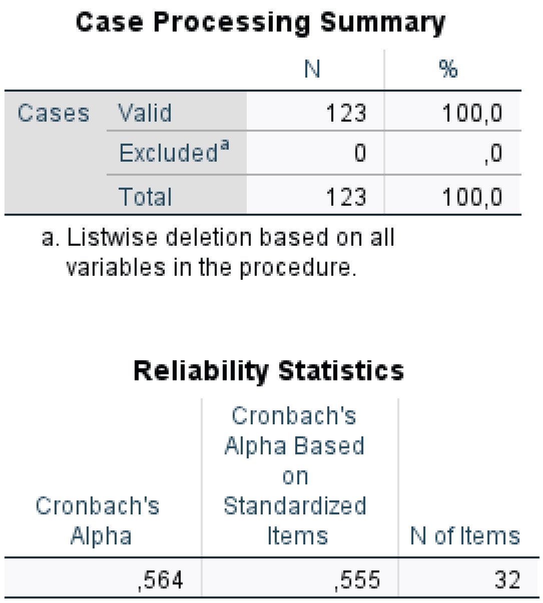
*Mark only one oval.*

| | 1 | 2 | 3 | 4 | 5 | |
|---|---|---|---|---|---|---|
| Strongly disagree | ○ | ○ | ○ | ○ | ○ | Strongly agree |

## Annex C. Users' experiences with deepfakes in Greece

The analysis of the Likert scale used in questionnaire was conducted with SPSS. At the same time Cronbach's alpha was used to measure the validity of the questionnaire. The Cronbach's reliability test which was performed on the dataset shows the internal consistency of the questionnaire and is a measurement of the questionnaire reliability [156] [157]. The Table C-1 describes the Reliability Statistics results which provides the Cronbach's alpha value.

Table C - 1. Reliability analysis in SPSS



**Case Processing Summary**

| | | N | % |
|---|---|---|---|
| Cases | Valid | 123 | 100,0 |
| | Excluded[a] | 0 | ,0 |
| | Total | 123 | 100,0 |

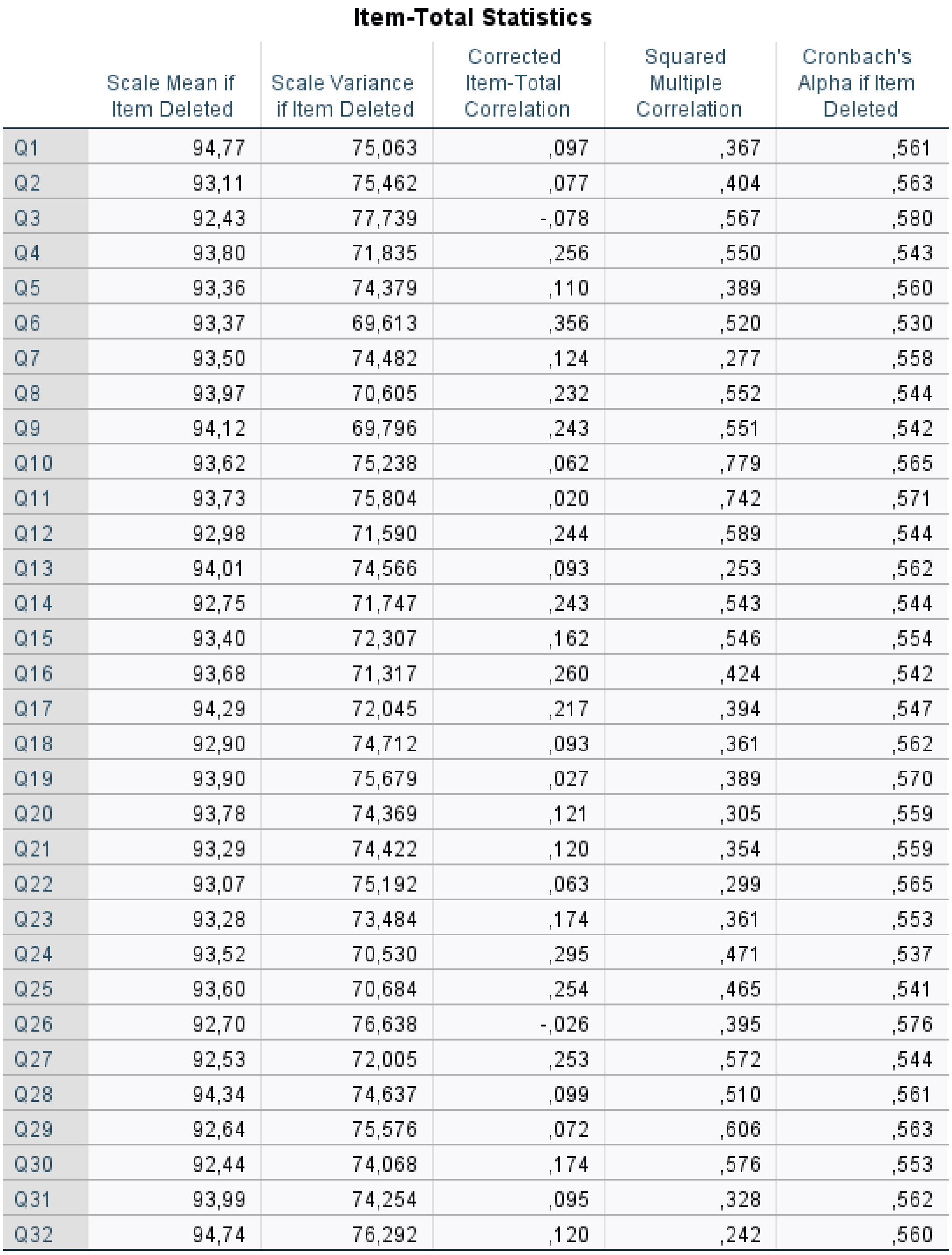
a. Listwise deletion based on all variables in the procedure.

**Reliability Statistics**

| Cronbach's Alpha | Cronbach's Alpha Based on Standardized Items | N of Items |
|---|---|---|
| ,564 | ,555 | 32 |

The alpha coefficient for the 32 items (32 statements in the questionnaire) is 0.564. According to the authors of "SPSS Explained" [158], this measurement is an acceptable Cronbach's alpha. At the same time, the Table C-2 shows how removing some items of the measuring instrument would impact the measurements of reliability and Cronbach's alpha. For instance, deletion of the items Q3 and Q27 would improve the result of Cronbach's alpha, which would reach 0.580.

Table C-  2. Item Total Statistics in SPSS



**Item-Total Statistics**

| | Scale Mean if Item Deleted | Scale Variance if Item Deleted | Corrected Item-Total Correlation | Squared Multiple Correlation | Cronbach's Alpha if Item Deleted |
|---|---|---|---|---|---|
| Q1 | 94,77 | 75,063 | ,097 | ,367 | ,561 |
| Q2 | 93,11 | 75,462 | ,077 | ,404 | ,563 |
| Q3 | 92,43 | 77,739 | -,078 | ,567 | ,580 |
| Q4 | 93,80 | 71,835 | ,256 | ,550 | ,543 |
| Q5 | 93,36 | 74,379 | ,110 | ,389 | ,560 |
| Q6 | 93,37 | 69,613 | ,356 | ,520 | ,530 |
| Q7 | 93,50 | 74,482 | ,124 | ,277 | ,558 |
| Q8 | 93,97 | 70,605 | ,232 | ,552 | ,544 |
| Q9 | 94,12 | 69,796 | ,243 | ,551 | ,542 |
| Q10 | 93,62 | 75,238 | ,062 | ,779 | ,565 |
| Q11 | 93,73 | 75,804 | ,020 | ,742 | ,571 |
| Q12 | 92,98 | 71,590 | ,244 | ,589 | ,544 |
| Q13 | 94,01 | 74,566 | ,093 | ,253 | ,562 |
| Q14 | 92,75 | 71,747 | ,243 | ,543 | ,544 |
| Q15 | 93,40 | 72,307 | ,162 | ,546 | ,554 |
| Q16 | 93,68 | 71,317 | ,260 | ,424 | ,542 |
| Q17 | 94,29 | 72,045 | ,217 | ,394 | ,547 |
| Q18 | 92,90 | 74,712 | ,093 | ,361 | ,562 |
| Q19 | 93,90 | 75,679 | ,027 | ,389 | ,570 |
| Q20 | 93,78 | 74,369 | ,121 | ,305 | ,559 |
| Q21 | 93,29 | 74,422 | ,120 | ,354 | ,559 |
| Q22 | 93,07 | 75,192 | ,063 | ,299 | ,565 |
| Q23 | 93,28 | 73,484 | ,174 | ,361 | ,553 |
| Q24 | 93,52 | 70,530 | ,295 | ,471 | ,537 |
| Q25 | 93,60 | 70,684 | ,254 | ,465 | ,541 |
| Q26 | 92,70 | 76,638 | -,026 | ,395 | ,576 |
| Q27 | 92,53 | 72,005 | ,253 | ,572 | ,544 |
| Q28 | 94,34 | 74,637 | ,099 | ,510 | ,561 |
| Q29 | 92,64 | 75,576 | ,072 | ,606 | ,563 |
| Q30 | 92,44 | 74,068 | ,174 | ,576 | ,553 |
| Q31 | 93,99 | 74,254 | ,095 | ,328 | ,562 |
| Q32 | 94,74 | 76,292 | ,120 | ,242 | ,560 |

The questionnaire included 32 statements to be rated by responders on the scale of 1 to 5. The analysis on the dataset of responses was conducted with the use of IBM SPSS Statistics and MS Office Excel graphs. The measurement of central tendencies of each statement will be done with the use of median and not the use of mean of the collected responses. The median is the suitable measurement to obtain the “central tendency” and average of the Likert scale responses [159] [160] [161]. Likert scale items do not produce continuous data, but ordinal. This means that Likert scale cannot yield mean values of 3.09 for example, which are produced by calculating the mean. Thus, the median which is "the point that separates the upper and lower halves of the data if we arrange them from largest to smallest" will be used instead of the mean [159].

The first statement which was addressed to the survey participants was "I trust everything posted online and/or on social media.". The Table C-3 shows the frequency of each statement as measured in a 5-point Likert scale. More than half of the asked users strongly disagree with the above statement. Around 30% disagree, while 15% individuals claimed to have a neutral stance. Only 5 individuals agreed or strongly agreed to be trustful towards the content posted online and/or on social media.

Table C-3. First statement response frequencies

| **1.I trust everything posted online and/or on social media.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 63 | 36 | 19 | 4 | 1 |
| **Percentages** | 51.2% | 29.3% | 15.4% | 3.3% | 0.8% |

At the same time, the Table C-4 shows that the maximum response to the first question was 5 and the minimum was 1, based on the 5-point Likert scale. The median is 1, which indicates that majority of the survey participants strongly disagreed with the first statement. They claim not to trust everything posted online and on social media platforms.

Table C-4. First statement measurements

| **Statement** | **1.I trust everything posted online and/or on social media** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |

| | |
|---|---|
| **Mean** | 1.73 |
| **Standard deviation** | 0.897 |
| **Median** | 1 |
| **Range** | 4 |
| **Total** | 123 |

The following Figure C-1 illustrates the frequencies and percentages allocated for each response to the first question.

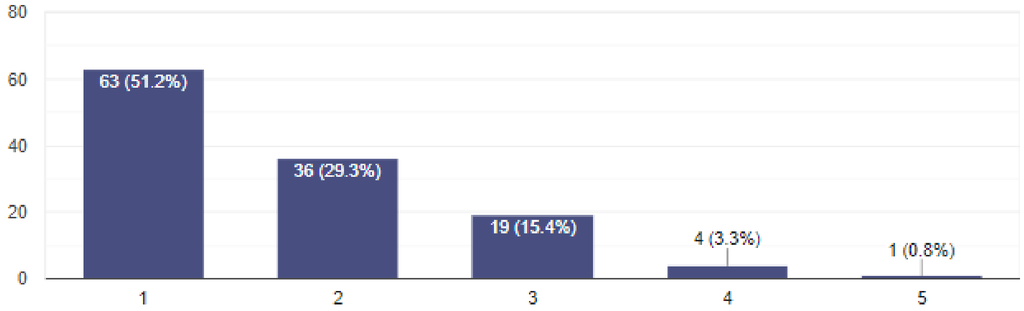


Figure C-1. First statement response percentages

The second statement in the questionnaire was "I find online and/or social media content as misleading and harmful for users." The most frequent response was "Neutral", which was counted 61 out 123 in total. Almost half of the responders did not indicate a clear stance towards the content on social media and online. Around 27% of the participants agreed that the content is harmful and misleading. The measured median of the collected responses stands at 3, as seen in the Table C-5. This shows that most responders are neutral.

Table C-5. Second statement response frequencies

| **2.I find online and/or on social media content as misleading** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| **and harmful for users.** |  |  |  |  |  |
| **Frequencies** | 1 | 13 | 61 | 33 | 15 |
| **Percentages** | 0.8% | 10.6% | 49.6% | 26.8% | 12.2% |

The median of the responses is calculated as 3 which shows that the central tendency of the responders is neutrality. This can be viewed in the above Table C-6.

Table C-6. Second statement measurements

| **Statement** | **2.I find online and/or social media content as misleading and harmful for users.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 3.39 |
| **Standard deviation** | 0.865 |
| **Median** | 3 |
| **Range** | 4 |
| **Total** | 123 |

The Figure C-2 puts the above data into a graphical representation.

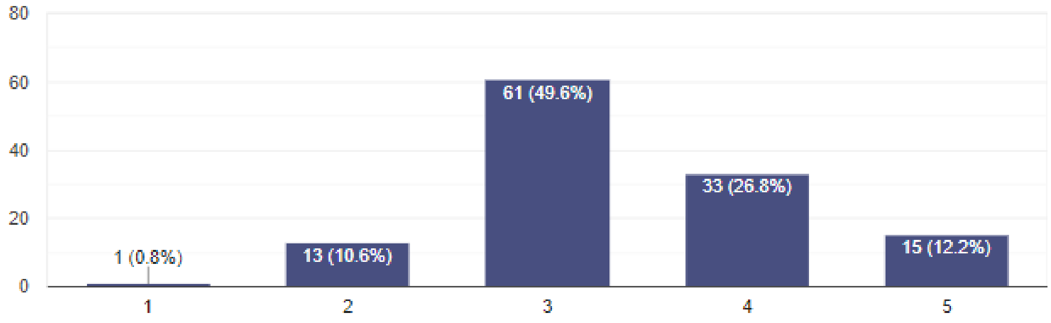


Figure C-2. Second statement response percentages

The third statement “I approach critically the content posted online and/or on social media” gathered the following responses, as described in the Table C-7.

Table C-7. Third statement response frequencies

| **3.I approach critically the content posted online and/or on social media.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 3 | 5 | 24 | 39 | 52 |
| **Percentages** | 2.4% | 4.1% | 19.5% | 31.7% | 42.3% |

Table C-8. Third statement measurements

| **Statement** | **3.I approach critically the content posted online and/or on social media.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 4.07 |
| **Standard deviation** | 1.001 |

|  |  |
|---|---|
| **Median** | 4 |
| **Range** | 4 |
| **Total** | 123 |

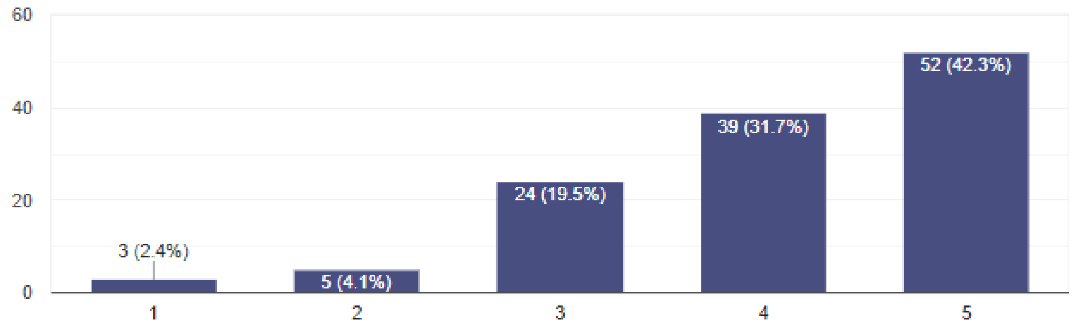


Figure C-3. Third statement response percentages

The Tables C-7 and C-8 and Figure C-3, show that most responders strongly agree to have a critical stance towards social media and online content. The median supports this view, as it stands at 4. Thus, the trend is towards agreement to the statement.

The next statement was “All deepfakes are malicious”. The results are tight. The Table C-9 shows that 43 responses were measured for the “Neutral” and 42 for “Disagree”. The median of the responses is 3, as described in the Table C-10. This indicates an inclination towards neutrality.

Table C-9. Fourth statement response frequencies

| **4.All deepfakes are malicious.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 13 | 42 | 43 | 18 | 7 |
| **Percentages** | 10.6% | 34.1% | 35% | 14.6% | 5.7% |

Table C-10. Fourth statement measurements

| **Statement** | **4.All deepfakes are malicious** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 2.17 |
| **Total** | 123 |

The Figure C-4 illustrates the responses collected for the fourth statement. The picture clearly shows that the users were not single minded and that 2-disagree and 3-neutral came close.

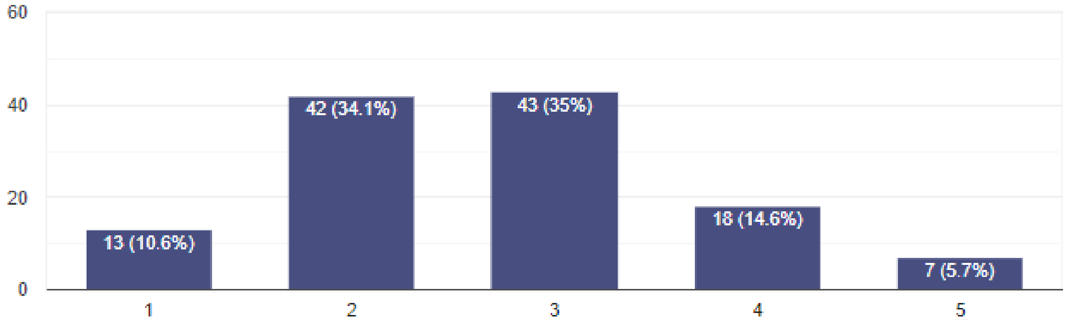


Figure C-4. Fourth statement response percentages

The fifth statement “Deepfakes benefit some spheres of life such as education and entertainment” collected the following responses, as viewed in the Tables C-11 and C-12, and depicted graphically in Figure C-5.

Table C-11. Fifth statement response frequencies

| **5.Deepfakes benefit some spheres of life such as** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **education and entertainment.** | | | | | |
| **Frequencies** | 7 | 25 | 44 | 37 | 10 |
| **Percentages** | 5.7% | 20.3% | 35.8% | 30.1% | 8.1% |

Table C-12. Fifth statement measurements

| **Statement** | **5.Deepfakes benefit some spheres of life such as education and entertainment.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 3.15 |
| **Standard deviation** | 1.022 |
| **Median** | 3 |
| **Range** | 4 |
| **Total** | 123 |

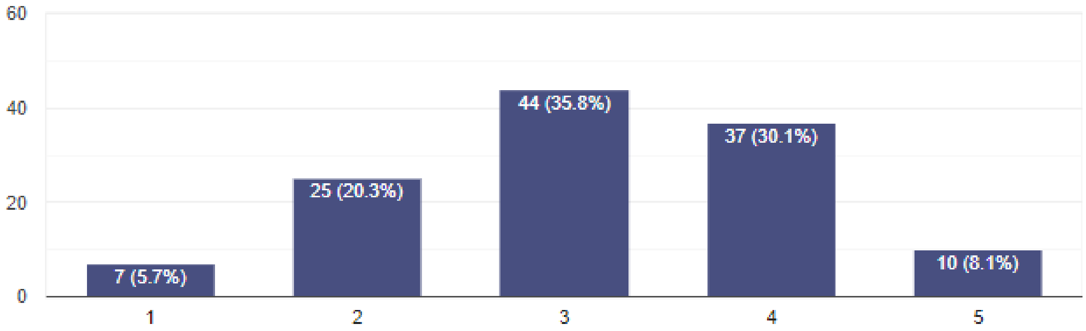


Figure C-5. Fifth statement response percentages

Most frequent is the response “Neutral”. Only a few responses behind stands the response “Agree”. Rest options – “Strongly disagree”, “Disagree” and “Strongly agree” were notably less frequent. The median of collected answers is 3. This means that the central tendency in this case is neutrality.

The attitudes towards the sixth statement “Most deepfakes are a threat to my cybersecurity” are collected in the Tables C-13 and C-14 and Figure C-6.

Table C-13. Sixth statement response frequencies

| **6. Most deepfakes are a threat to my cybersecurity.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 9 | 28 | 36 | 38 | 12 |
| **Percentages** | 7.3% | 22.8% | 29.3% | 30.9% | 9.8% |

Table C-14. Sixth statement measurements

| **Statement** | **6. Most deepfakes are a threat to my cybersecurity.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 3.13 |
| **Standard deviation** | 1.101 |
| **Median** | 3 |
| **Range** | 4 |
| **Total** | 123 |

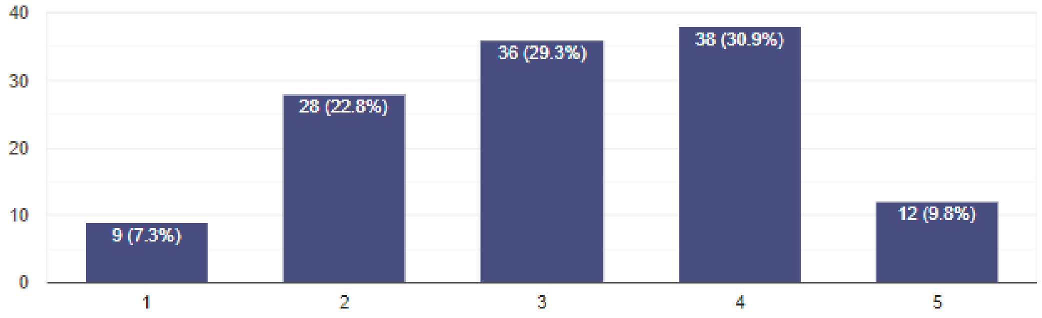


Figure C-6. Sixth statement response percentages.

The responders were not single minded. The choice “Neutral” gathered 36 responses, while the option “Agree” collected 2 more responses. There is no emphatic attitude towards the declaration that most deepfakes pose a threat to users’ cybersecurity. The central tendency described by the median is neutrality and 3.

Moving on to the seventh statement, most responders are neutral towards the statement “I feel confident in detecting deepfakes”. The Tables C-15 and C-16 as well as the Figure C-7 illustrate the distribution of the responses.

Table C-15. Seventh statement response frequencies

| **7. I feel confident in detecting deepfakes.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 7 | 27 | 52 | 32 | 5 |
| **Percentages** | 5.7% | 22% | 42.3% | 26% | 4.1% |

Table C-16. Seventh statement measurements

| **Statement** | **7. I feel confident in detecting deepfakes.** |
|---|---|

| | |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 3.01 |
| **Standard deviation** | 0.936 |
| **Median** | 3 |
| **Range** | 4 |
| **Total** | 123 |

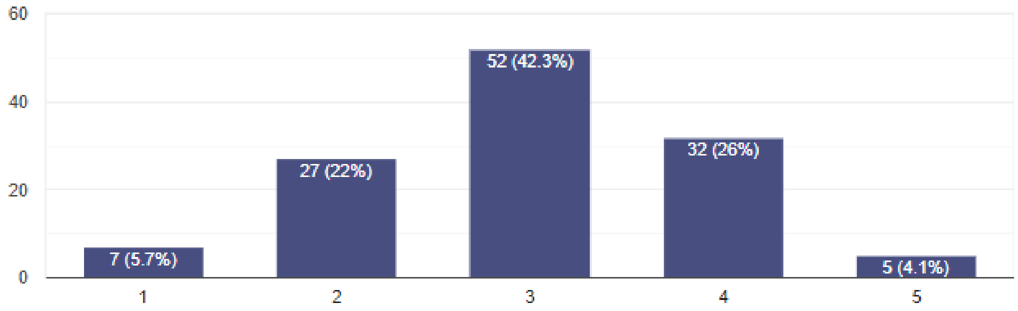


Figure C-7. Seventh statement response percentages

42.3% of responders have a neutral stance towards their confidence and ability of detecting deepfakes. 26% of the participants agree that they are confident in detecting the AI manipulated content, while 22% claims not to agree with this statement. The least frequent were the responses of “Strongly agree” (4.1%) and “Strongly disagree” (5.7%). The calculated median is 3. Thus, the central tendency forms towards neutrality.

Next statement was “I have experienced a phishing attempt with the use of a deepfake”. The responses to this statement show a more determined distribution of responses. Specifically, Table C-17 depicts that 36 responders strongly disagree to having fallen victim to a deepfake phishing attack. 28 responders disagree to this statement. Collectively, 64 out of 123 participants were negative to the eighth statement. Equal distribution can be viewed in the responses of “Neutral” and “Agree”, as each collected 25 responses. Finally, as only 9 individuals strongly agree that they have experienced a phishing attempt with the use of a deepfake. In total, 34 out of the 123 responders claim to be a targeted by a phishing deepfake attack. The general inclination is 2 and “Disagree”, as measured by the median (see Table C-18).

Table C-17. Eighth statement response frequencies

| 8. I have experienced a phishing attempt with the use of a deepfake. | Strongly disagree (1) | Disagree (2) | Neutral (3) | Agree (4) | Strongly agree (5) |
|---|---|---|---|---|---|
| **Frequencies** | 36 | 28 | 25 | 25 | 9 |
| **Percentages** | 29.3% | 22.8% | 20.3% | 20.3% | 7.3% |

Table C-18. Eighth statement measurements

| Statement | 8. I have experienced a phishing attempt with the use of a deepfake. |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 2.54 |
| **Standard deviation** | 1.301 |
| **Median** | 2 |
| **Range** | 4 |
| **Total** | 123 |

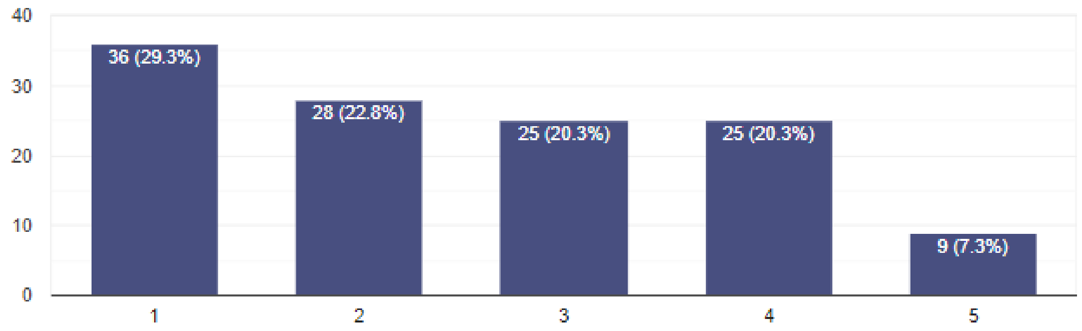


Figure C-8. Eighth statement response percentages

The ninth statement aims to measure the exposure of the survey participants to frauds with the use of deepfakes. Specifically, the responses to the statement "I have experienced an online fraud with the use of a deepfake" are gathered in the Tables C-19 and C-20 and Figure C-9.

Table C-19. Ninth statement response frequencies

| **9. I have experienced an online fraud with the use of a deepfake.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 48 | 23 | 21 | 19 | 12 |
| **Percentages** | 39% | 18.7% | 17.1% | 15.4% | 9.8% |

Table C-20. Ninth statement measurements

| **Statement** | **9. I have experienced an online fraud with the use of a deepfake.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |

| | |
|---|---|
| **Mean** | 238 |
| **Standard deviation** | 1.388 |
| **Median** | 2 |
| **Range** | 4 |
| **Total** | 123 |

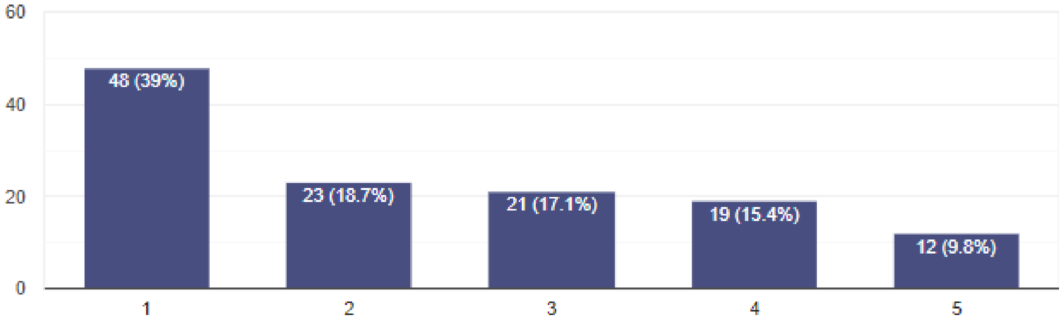


Figure C-9. Ninth statement response percentages

The frequencies denote a decisive and emphatic attitude of the survey responders. In details, 48 out of 123 participants strongly disagree with having experienced an online fraud involving deepfakes. Additionally, 23 individuals disagree to this statement. All in all, 71 of 123 people have not experienced this cyber threat. The median 2 supports the central tendency of disagreement in this scale.

Next participants were confronted with three statements regarding their level of security. Specifically, the tenth statement was “In relation to deepfakes, I feel secure browsing on the Internet”. The eleventh statement was “In relation to deepfakes, I feel secure while using social media”. The twelfth statement was “In relation to deepfakes, I am worried about my cybersecurity and privacy online”.

The Tables C-21 and C-22 cover the frequencies and percentages of each option in the Likert scale. Specifically, most responders are neutral towards the statement “In relation to deepfakes, I feel secure browsing the Internet”. The median of 3 confirms this central tendency.

Table C-21. Tenth statement response frequencies

| **10. In relation to deepfakes, I feel secure browsing on the Internet.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 10 | 34 | 46 | 26 | 7 |
| **Percentages** | 8.1% | 27.6% | 37.4% | 21.1% | 5.7% |

Table C-22. Tenth statement measurements

| **Statement** | **10. In relation to deepfakes, I feel secure browsing on the Internet** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 2.89 |
| **Standard deviation** | 1.018 |
| **Median** | 3 |
| **Range** | 4 |
| **Total** | 123 |

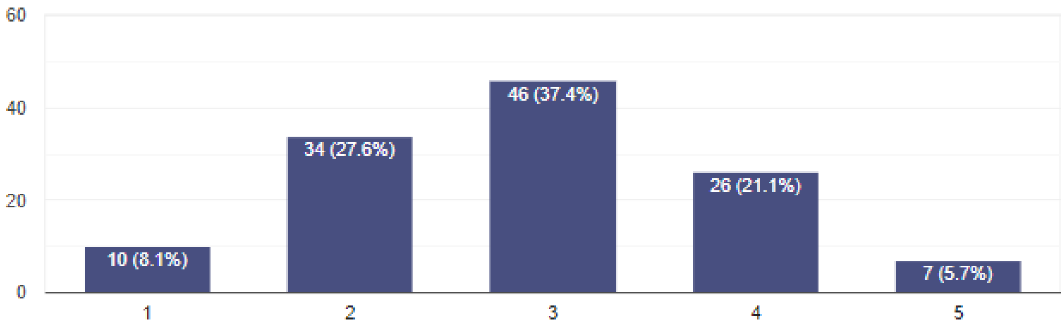


Figure C-10. Tenth statement response percentages

In contrast to the above result of neutrality, most users disagreed with the eleventh statement "In relation to deepfakes, I feel secure while using social media". A slight less, have a neutral attitude towards this declaration. The median is calculated as 3, which shows that the central tendency is neutrality. This comes from the fact that the frequencies of "Neutral" (3) and "Agree" (4) were collectively higher than the frequencies expressing disagreement. Thus, the inclination is towards to the right columns of the below Table C-23.

Table C-23. Eleventh statement response frequencies

| **11. In relation to deepfakes, I feel secure while using social media.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 14 | 40 | 36 | 26 | 7 |
| **Percentages** | 11.4% | 32.5% | 29.3% | 21.1% | 5.7% |

Table C-24. Eleventh statement measurements

| **Statement** | **11. In relation to deepfakes, I feel secure while using social media.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 2.77 |
| **Standard deviation** | 1.085 |
| **Median** | 3 |
| **Range** | 4 |
| **Total** | 123 |

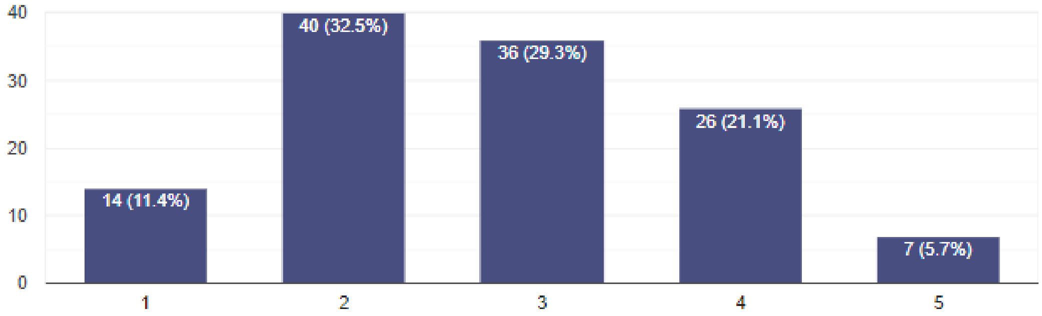


Figure C-11. Eleventh statement response percentages

Finally, the twelfth statement focuses on the deepfakes threat to cybersecurity and privacy of users as perceived by the survey participants. The most frequent response was “Agree”, since it was counted 38 times. Four less gathered the “Neutral” option in the scale, followed by “Strongly agree” with 27 responses. All results can be viewed in the Table C-25. The central tendency is 4 which corresponds to “Agree” with the statement.

Table C-25. Twelfth statement response frequencies

| **12. In relation to deepfakes, I am worried about my cybersecurity and privacy online.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 4 | 20 | 34 | 38 | 27 |
| **Percentages** | 3.3% | 16.3% | 27.6% | 30.9% | 22% |

Table C-26. Twelfth statement measurements

| **Statement** | **12. In relation to deepfakes, I am worried about my cybersecurity and privacy online.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 3.52 |
| **Standard deviation** | 1.104 |
| **Median** | 4 |
| **Range** | 4 |
| **Total** | 123 |

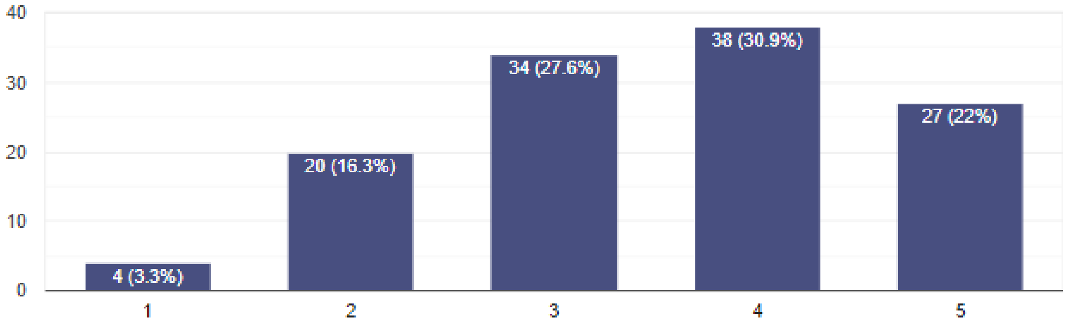


Figure C-12. Twelfth statement response percentages

The thirteenth, fourteenth and fifteenth statements aim to describe the behaviour of users when they come across a deepfake.

The prevailing response to the statement “When I come across a potential deepfake, I accept that I am helpless, and I cannot trust anything online” is “Disagree”, which means that users support a more proactive stance towards deepfakes. The second most frequent response is “Neutral” with count 38. 21 individuals were in strong disagreement with being helpless over deepfakes online. Only 20 out of 123 responders were agreeing or strongly agreeing with being in helpless position in regarding to deepfakes. The Table C-27 describes some additional measurements regarding the responses. The median is calculated as 2. “Disagree” is the prevailing trend in this statement. Finally, the Figure C-13 constitutes a graphical representation of the percentages of each response.

Table C-27. Thirteenth statement response frequencies

| **13. When I come across a potential deepfake, I accept that I am helpless, and I cannot trust anything online.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 21 | 45 | 38 | 13 | 6 |
| **Percentages** | 17.1% | 36.6% | 30.9% | 10.6% | 4.9% |

Table C-28. Thirteenth statement measurements

| **Statement** | **13. When I come across a potential deepfake, I accept that I am helpless, and I cannot trust anything online.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 2.50 |
| **Standard deviation** | 1.051 |
| **Median** | 2 |
| **Range** | 4 |
| **Total** | 123 |

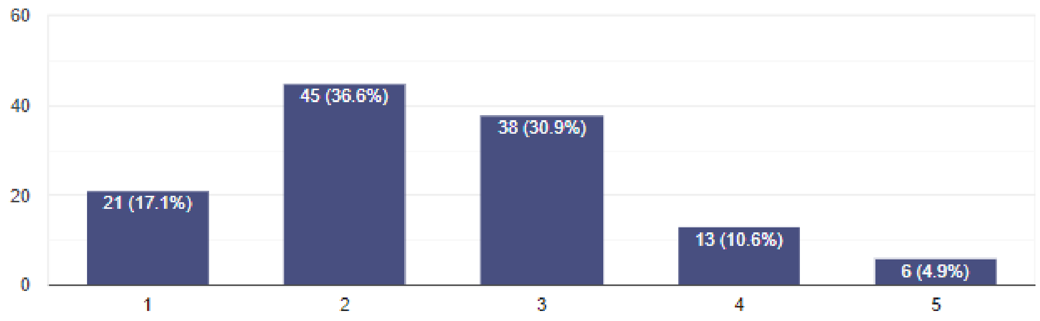


Figure C-13. Thirteenth statement response percentages

The proactive attitude towards deepfakes can be also viewed by the responses to the fourteenth statement, “When I come across a potential deepfake, I compare this media with other reports online”. The following Tables C-29 and C-30 and Figure C-14 display the results.

Table C-29. Fourteenth statement response frequencies

| **14. When I come across a potential deepfake, I compare this media with other reports online.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 6 | 9 | 27 | 48 | 33 |
| **Percentages** | 4.9% | 7.3% | 22% | 39% | 26.8% |

Table C-30. Fourteenth statement measurements

| **Statement** | **14. When I come across a potential deepfake, I compare this media with other reports online.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 3.76 |
| **Standard deviation** | 1.081 |
| **Median** | 4 |
| **Range** | 4 |
| **Total** | 123 |

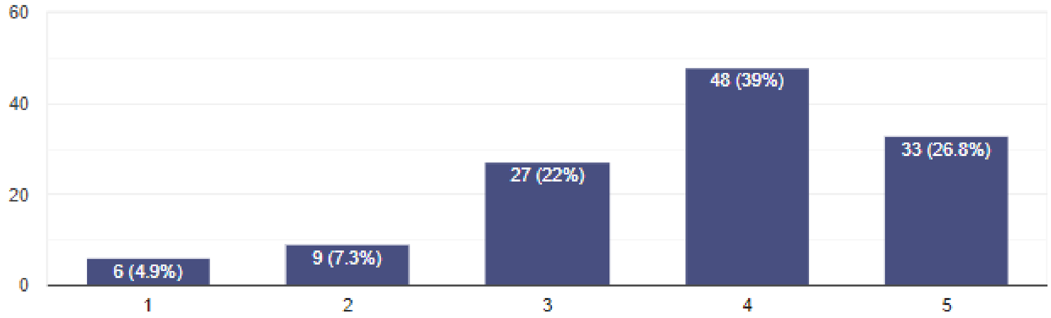


Figure C-14. Fourteenth statement response percentages

Most responders tend to “Agree” or “Strongly agree” that they compare deepfakes and other sources to check information. The median of 4 confirms this inclination.

The following Tables C-31 and C-32 and Figure C-15 examine the distribution of responses to the statement “When I come across a potential deepfake, I do a reverse image search to see if there are any existing photos or videos that might have been used as a base for a deepfake”.

Table C-31. Fifteenth statement response frequencies

| **15. When I come across a potential deepfake, I do a reverse image search to see if there are any existing photos or videos that might have been used as a base for a deepfake.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 18 | 19 | 36 | 32 | 18 |
| **Percentages** | 14.6% | 15.4% | 29.3% | 26% | 14.6% |

Table C-32. Fifteenth statement measurements

| **Statement** | **15. When I come across a potential deepfake, I do a reverse image search to see if there are any existing photos or videos that might have been used as a base for a deepfake.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 3.11 |
| **Standard deviation** | 1.260 |
| **Median** | 3 |
| **Range** | 4 |

| **Total** | 123 |
|---|---|

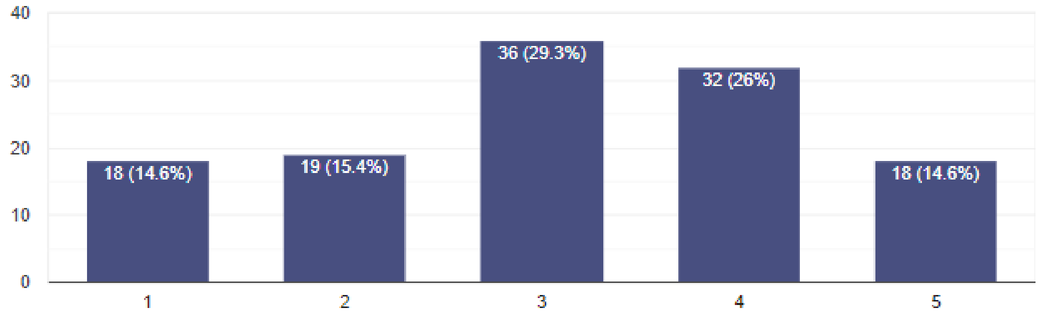


Figure C-15. Fifteenth statement response percentages

The reverse search of images received most “Neutral” responses. 36 responders made this selection. 32 responders agreed with doing reverse image search to detect a deepfake. The other responses of “Strongly disagree”, “Disagree” and “Strongly agree” collected similar rates of responses. It was around 15% for each. In terms of central tendency, the neutrality can be confirmed by the median calculated as 3.

The next issue address is related to the responsibility of handling the threat of deepfakes. Specifically, the users were asked about the actors who should take care of the deepfakes spread and their level of confidence in the efforts of this actor.  First, the questionnaire asked the users to rate their level of agreement with the statement that the Greek government should be the main actor in fight against malicious deepfakes. The responses are shown in the following Table C-33 and C-34. The Figure C-16 illustrates the portion of each response.

Table C-33. Sixteenth statement response frequencies

| **16. The Greek government is the main responsible for stopping the spread of deepfakes** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **online for my cybersecurity.** | | | | | |
| **Frequencies** | 15 | 35 | 37 | 29 | 7 |
| **Percentages** | 12.2% | 28.5% | 30.1% | 29% | 5.7% |

Table C-34. Sixteenth statement measurements

| **Statement** | **16. The Greek government is the main responsible for stopping the spread of deepfakes online for my cybersecurity.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 2.82 |
| **Standard deviation** | 1.102 |
| **Median** | 3 |
| **Range** | 4 |
| **Total** | 123 |

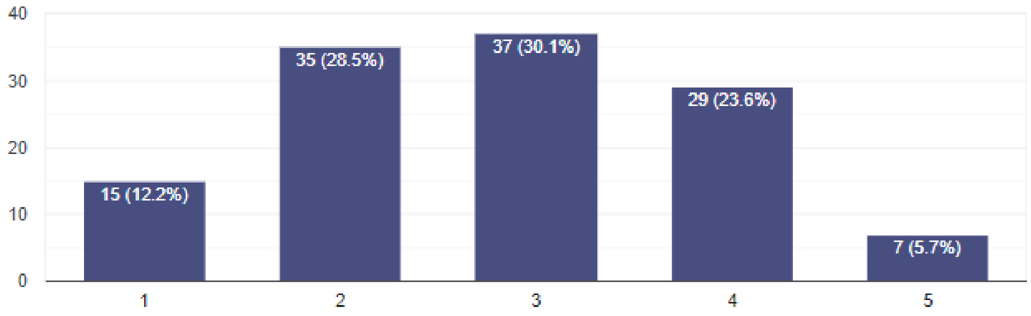


Figure C-16. Sixteenth statement response percentages.

Most responses were "Neutral", as this option was selected by 37 responders. 2 less responses were received by "Disagree". 15 individuals strongly disagree that Greek government is solely responsible to control the spread of the deepfakes online. 29

participants agree that the Greek government is the main responsible for countering this threat. Finally,7 users claimed to strongly agree with this statement. The median of 3 indicates that the central tendency towards this statement is neutrality.

When users were asked to rate their level of confidence towards the efforts of Greek government to resolve the issue of malicious deepfakes, the users incline towards lack of confidence. All details can be viewed in the Tables C-35 and C-36 and graphically in the Figure C-17.

Table C-35. Seventeenth statement response frequencies

| **17. I am confident in the government's efforts to combat the problem of malicious deepfakes online.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 41 | 34 | 34 | 9 | 5 |
| **Percentages** | 33.3% | 27.6% | 27.6% | 7.3% | 4.1% |

Table C-36. Seventeenth statement measurements

| **Statement** | **17. I am confident in the government's efforts to combat the problem of malicious deepfakes online.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 2.21 |
| **Standard deviation** | 1.111 |

| | |
|---|---|
| **Median** | 2 |
| **Range** | 4 |
| **Total** | 123 |

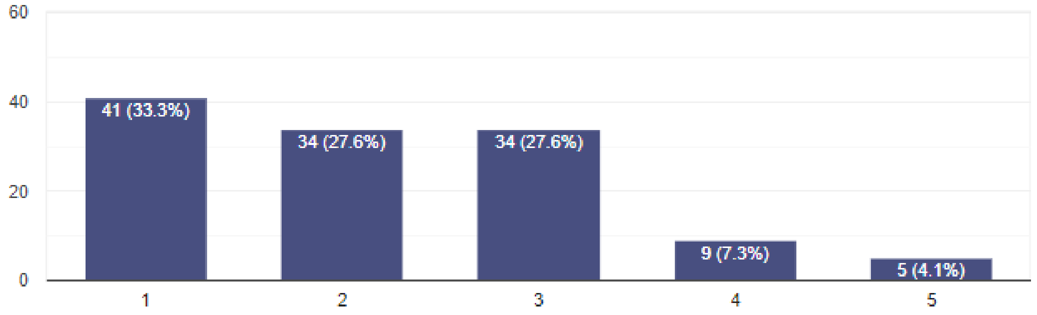


Figure C-17. Seventeenth statement response percentages

The most frequent answer was “Strongly disagree”, which was selected 41 times. Equally 34 responses collected the options of “Disagree” and “Neutral”. Only 14 individuals agreed or strongly agreed with the being confident of the Greek government efforts in the field of malicious deepfakes. The median of 2 clearly shows that “Disagreement” is the representative response to the 17th statement.

Next, the users were asked about the responsibility of the social media companies in the fight against deepfakes. Their responses can be viewed in the Tables C-37 and C-38 and Figure C-18.

Table C-37. Eighteenth statement response frequencies

| **18. Social media platforms are the main responsible for stopping the spread of deepfakes** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **online for my cybersecurity.** | | | | | |
| **Frequencies** | 4 | 13 | 34 | 49 | 23 |
| **Percentages** | 3.3% | 10.6% | 27.6% | 39.8% | 18.7% |

Table C-38. Eighteenth statement measurements.

| **Statement** | **18. Social media platforms are the main responsible for stopping the spread of deepfakes online for my cybersecurity.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 3.60 |
| **Standard deviation** | 1.014 |
| **Median** | 4 |
| **Range** | 4 |
| **Total** | 123 |

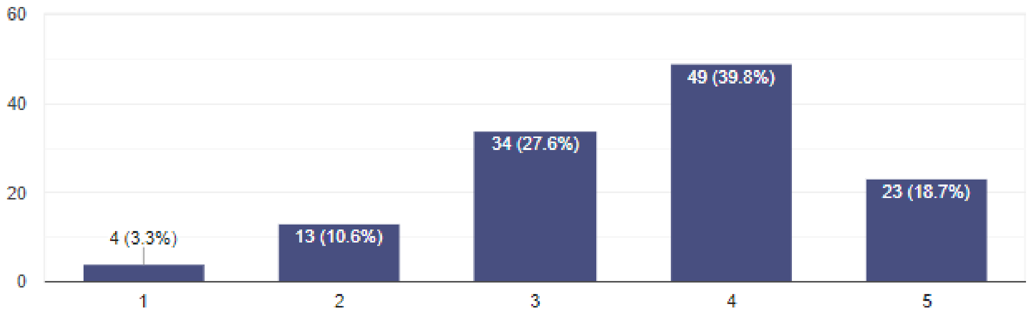


Figure C-18. Eighteenth statement response percentages

Most users agree that the social media companies have the main responsibility to control the spread of malicious deepfakes online. Almost 40% of all participants selected this option. Circa 28% of collected responses disclose neutral stance towards this statement. Almost 19% strongly agree. The least popular were the options “Disagree” (10.6%) and “Strongly disagree” (3.3%). The tendency is 4 which represents an agreement with the social media with main responsible to combat malicious deepfakes.

Regarding the level of confidence in the AI technologies used by social media companies for deepfakes detection, the users responded as shown in the Tables C-39 and C-40 and Figure C-19.

Table C-39. Ninetieth statement response frequencies

| **19. Social media companies such as Facebook uses AI-based detection tools to prevent the spread of deepfakes on the platform. I have confidence in the technologies used by Facebook to protect its users from deepfakes.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 23 | 32 | 44 | 19 | 5 |
| **Percentages** | 18.7% | 26% | 35.8% | 15.4% | 4.1% |

Table C-40. Nineteenth statement measurements

| **Statement** | **19. Social media companies such as Facebook uses AI-based detection tools to prevent the spread of deepfakes on the platform. I have confidence in the technologies used by Facebook to protect its users from deepfakes.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 2.60 |
| **Standard deviation** | 1.084 |
| **Median** | 3 |
| **Range** | 4 |
| **Total** | 123 |

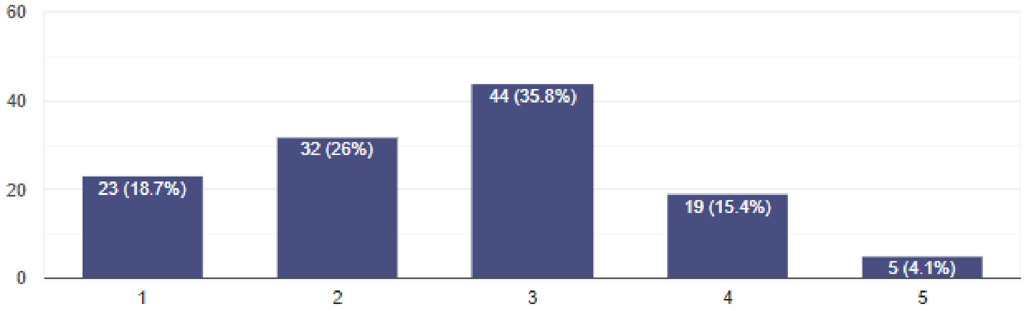


Figure C-19. Nineteenth statement response percentages.

Most users have a neutral stance towards the effectiveness of detection technologies of the social media platforms. 44 responses of "Neutral" were recorded. 32 users disagree with the 19th statement and 23 strongly disagree. 19 individuals do agree that the AI technologies are effective in the fight against deepfakes, while 5 users strongly agree with this statement. The calculated median of 3 supports that the prevailing tendency is neutrality. The users are not single-minded and confident in the technologies used by Facebook to prevent the spread of malicious deepfakes online.

Finally, the users were asked about whether the non-governmental and educational organisations hold the main responsibility for protecting users from malicious deepfakes.

Table C-41. Twentieth statement response frequencies

| **20. Non-governmental and educational organisations are the main responsible to protect users from deepfakes.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 12 | 39 | 48 | 19 | 5 |
| **Percentages** | 9.8% | 31.7% | 39% | 15.4% | 4.1% |

Table C-42. Twentieth statement measurements

| **Statement** | **20. Non-governmental and educational organisations are the main responsible to protect users from deepfakes.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 2.72 |
| **Standard deviation** | 0.987 |
| **Median** | 3 |
| **Range** | 4 |
| **Total** | 123 |

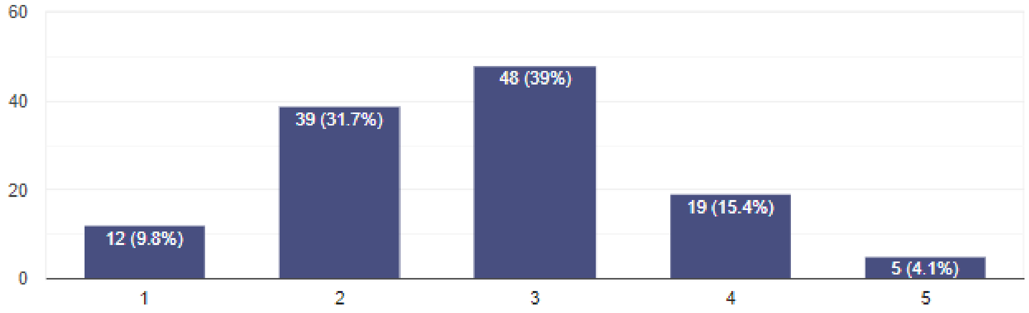


Figure C-20. Twentieth statement response percentages

As the above Tables C-41 and C-42 and Figure C-20 above show, the most frequent response recorded was “Neutral”. 48 individuals chose this option. Neutrality is the central tendency of all responses. The second most popular response was “Disagree”. 39 responders do not believe that the non-governmental a/or educational institutions are the main responsible for handling the threat of deepfakes. On the other hand, 19 responders advocate for the “Agreement” with the leading role of non-governmental and educational organisations in the fight against deepfakes.

The questionnaire included an example of a non-governmental organisation which contributes to the fight against malicious deepfakes on Facebook. Specifically, the users were asked to rate their confidence level in the efforts of Ellinika Hoaxes to spot and limit the spread of deepfakes. Most of the responders (60 out of 123) are neutral. 23 individuals feel confident and 11 feel strongly confident. Only 12 individuals were not confident with the Ellinika Hoaxes efforts, while 8 strongly disagree with being confident. The neutrality, expressed as the option 3 in the Likert scale, is the central tendency in the scale. All the above can be viewed in the Tables C-43 and C-44 as well as graphically in the Figure C-21.

Table C-43. Twenty-first statement response frequencies

| **21. Ellinika Hoaxes is an online Greek community** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **where users exchange knowledge and insights, and they collaborate to spot fake news and counter its spread. I am confident in their efforts to limit the spread of deepfakes online.** | | | | | |
| **Frequencies** | 8 | 12 | 60 | 32 | 11 |
| **Percentages** | 6.5% | 9.8% | 48.8% | 26% | 8.9% |

Table C-44. Twenty-first statement measurements

| **Statement** | **21. Ellinika Hoaxes is an online Greek community where users exchange knowledge and insights, and they collaborate to spot fake news and counter its spread. I am confident in their efforts to limit the spread of deepfakes online.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 3.21 |
| **Standard deviation** | 0.969 |

| | |
|---|---|
| **Median** | 3 |
| **Range** | 4 |
| **Total** | 123 |

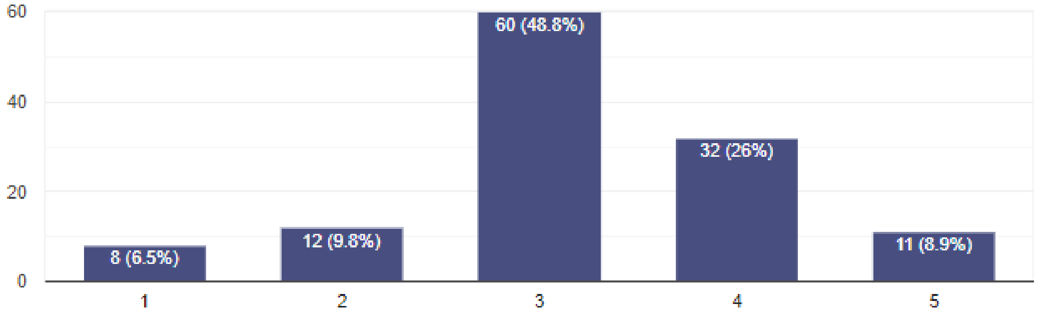


Figure C-21. Twenty-first statement response percentages

The next statement which was addressed to the questionnaire participants was "Technology-based methods are the only reliable and effective tool to combat deepfakes". 40 out of 123 participants were neutral towards this statement. 36 agreed and 22 strongly agreed with the above declaration. 24 out of 123 responders disagreed that technology provides the only effective and reliable method against malicious deepfakes. Only 1 person strongly disagreed. All analysis can be viewed in the Tables C-45 and C-46 and Figure C-22.

Table C-45. Twenty-second statement response frequencies

| **22. Technology-based methods are the only reliable and effective tool to combat deepfakes.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 1 | 24 | 40 | 36 | 22 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| **Percentages** | 0.8% | 19.5% | 32.5% | 29.3% | 17.9% |

Table C-46. Twenty-second statement measurements

| **Statement** | **22. Technology-based methods are the only reliable and effective tool to combat deepfakes.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 3.44 |
| **Standard deviation** | 1.025 |
| **Median** | 3 |
| **Range** | 4 |
| **Total** | 123 |

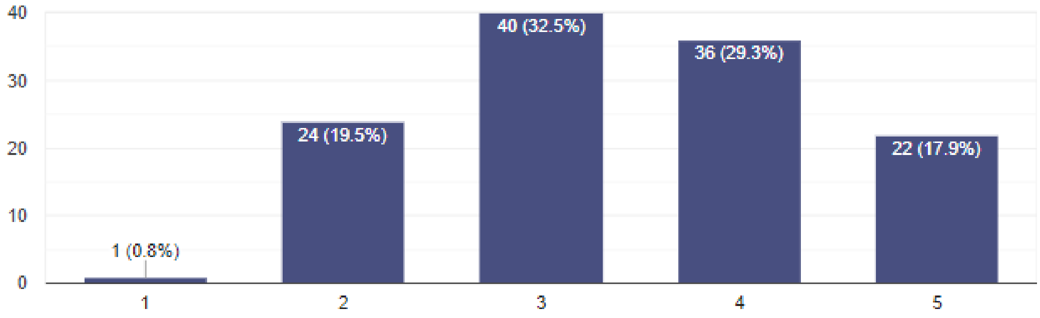


Figure C-22. Twenty-second statement response percentages.

The median is 3 which shows that the tendency of the collected responses is neutrality.

The users were asked to rate their agreement with the statement that user education is the key for combating deepfakes. Specifically, the responses were spread out as the Tables C-47 and C-48 and Figure C-23 depict.

Table C-47. Twenty-third statement response frequencies

| **23. User education and computer literacy are enough to deal with the threat of deepfakes and raise cybersecurity.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 4 | 25 | 45 | 38 | 11 |
| **Percentages** | 3.3% | 20.3% | 36.6% | 30.9% | 8.9% |

Table C-48. Twenty-third statement measurements

| **Statement** | **23. User education and computer literacy are enough to deal with the threat of deepfakes and raise cybersecurity.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 3.22 |
| **Standard deviation** | 0.980 |
| **Median** | 3 |
| **Range** | 4 |
| **Total** | 123 |

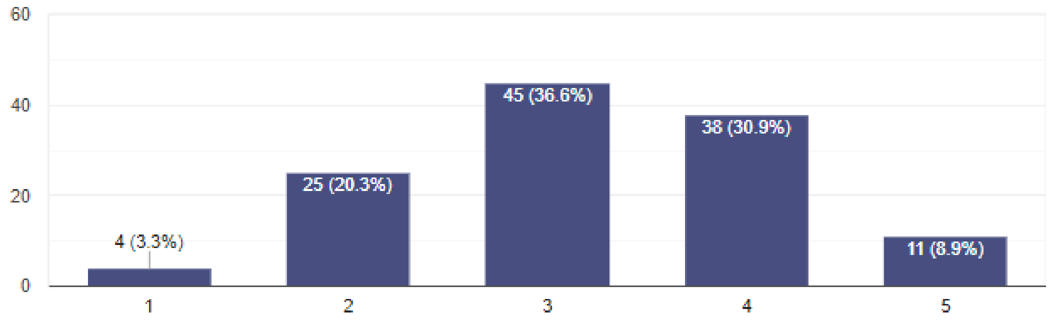


Figure C-23. Twenty-third statement response percentages

The prevailing trend is neutrality, since 45 out of 123 responders supported the option "Neutral". 38 individuals claimed to agree that education and computer skills are sufficient to combat the threat of deepfakes.  25 responses were collected for the option "Disagree". The least frequent were the responses of "Strongly agree" (11 responses) and "Strongly disagree" (4 responses).

Most users were also neutral towards the statement "Only legislation can stop the spread of malicious deepfakes online." A closer look on the Tables C-49 and C-50 and Figure C-24 reveals that circa 31% of the questionnaire responders are neutral, while almost 29% disagrees with the above statement. 22% claimed to agree that the only path towards fight against deepfakes is legislation. Finally, the extreme responses of "Strongly agree" and Strongly disagree" were the least popular among the responders. They gathered 13 and 11 responses respectively.

Table C-49. Twenty-fourth statement response frequencies

| **24. Only legislation can stop the spread of malicious deepfakes online.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 10 | 35 | 38 | 27 | 13 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| **Percentages** | 8.1% | 28.5% | 30.9% | 22% | 10.6% |

Table C-50. Twenty-fourth statement measurements

| **Statement** | **24. Only legislation can stop the spread of malicious deepfakes online.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 2.98 |
| **Standard deviation** | 1.123 |
| **Median** | 3 |
| **Range** | 4 |
| **Total** | 123 |

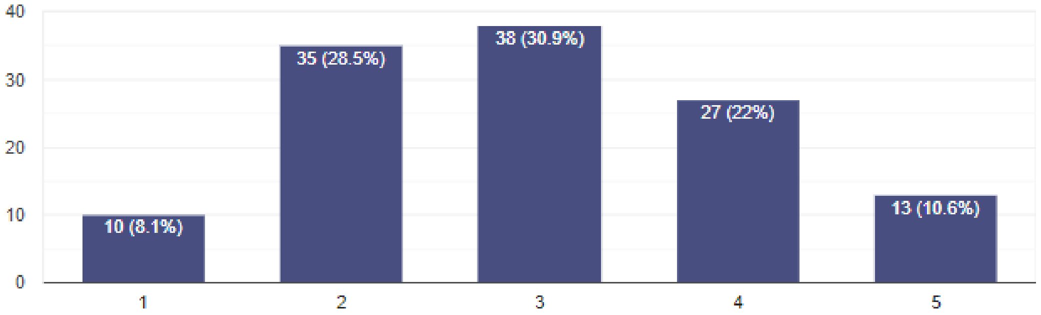


Figure C-24. Twenty-fourth statement response percentages

Going in depth into the legislation and relation to deepfakes, the users were asked to rate their level of agreement with the statement “All deepfakes should be banned.” Most responders expressed their disagreement with this statement. The details on responding amounts can be viewed in the Tables C-51 and C-52 and Figure C-25.

Table C-51. Twenty-fifth statement response frequencies

| 25. All deepfakes should be banned. | Strongly disagree (1) | Disagree (2) | Neutral (3) | Agree (4) | Strongly agree (5) |
|---|---|---|---|---|---|
| **Frequencies** | 15 | 37 | 31 | 25 | 15 |
| **Percentages** | 12.2% | 30.1% | 25.2% | 20.3% | 12.2% |

Table C-52. Twenty-fifth statement measurements

| Statement | 25. All deepfakes should be banned. |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 2.90 |
| **Standard deviation** | 1.217 |
| **Median** | 2 |
| **Range** | 4 |
| **Total** | 123 |

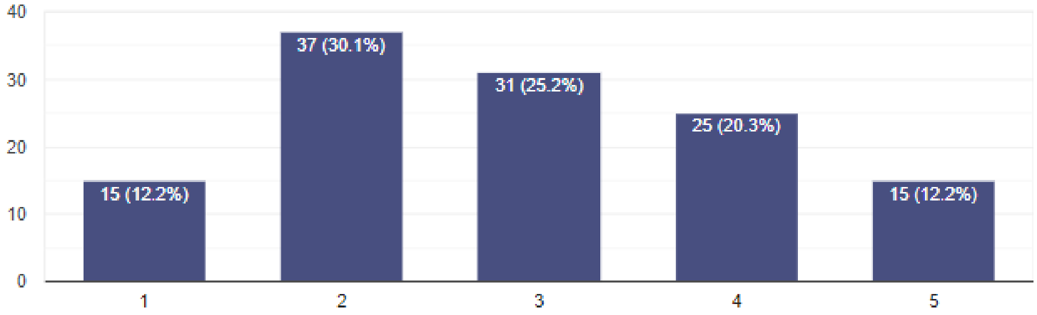


Figure C-25. Twenty-fifth statement response percentages

37 individuals concluded to agree that legislation should not ban all deepfakes. This is also expressed with the central tendency of the median 2.

The final statement regarding the relation of deepfakes and legislation is “Legislation should not ban all deepfakes. Laws should regulate the use of deepfakes to prohibit their use for illicit purposes.” The users were clear on their stance towards this statement. The Tables C-53 and C-54 and Figure C-26 show that 48 out of 123 agreed with the above declaration and 37 strongly agreed. This corresponds to 39 % and 30.1% respectively. Thus, the central tendency expressed by the median is calculated as 4 and “Agreement” in particular. The other options of “Neutral”, “Disagree” and “Strongly disagree” have much lower frequency rates.

Table C-53. Twenty-sixth statement response frequencies

| **26. Legislation should not ban all deepfakes. Laws should regulate the use of deepfakes to prohibit their use for illicit purposes.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 5 | 13 | 20 | 48 | 37 |
| **Percentages** | 4.1% | 10.6% | 16.3% | 39% | 30.1% |

Table C-54. Twenty-sixth statement measurements

| **Statement** | **26. Legislation should not ban all deepfakes. Laws should regulate the use of deepfakes to prohibit their use for illicit purposes.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |

| | |
|---|---|
| **Mean** | 3.80 |
| **Standard deviation** | 1.106 |
| **Median** | 4 |
| **Range** | 4 |
| **Total** | 123 |

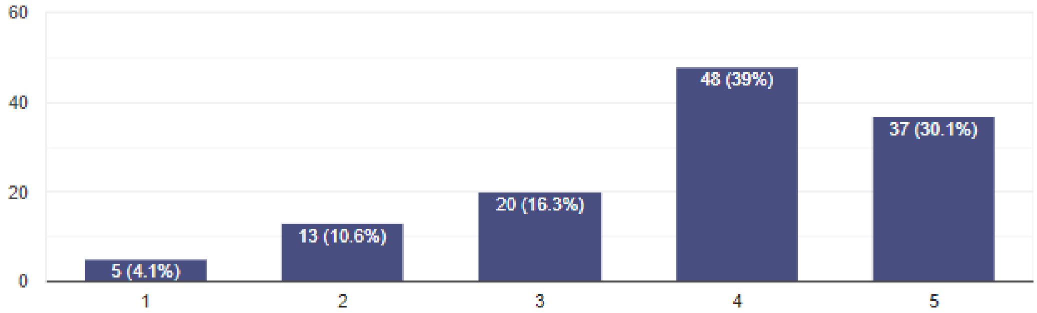


Figure C-26. Twenty-sixth statement response percentages

The questionnaire addressed the issue of collective efforts and cooperation in the fight against deepfakes. When asked if the European Union should be involved in the efforts of Greece to mitigate the threat of the deepfakes online, most users have a clear vision of the necessity of this joint action. 54 out of 123 responders agreed with the above declaration, while 41 strongly agreed. The rest available options did not have significant popularity. All details are described in the Tables C-55 and C-56 and Figure C-27.

Table C-55. Twenty-seventh statement response frequencies

| **27. The European Union should be involved in the efforts of Greece to** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **mitigate the threat of the deepfakes online.** | | | | | |
| **Frequencies** | 4 | 8 | 16 | 54 | 41 |
| **Percentages** | 3.3% | 6.5% | 13% | 43.9% | 33.3% |

Table C-56. Twenty-seventh statement measurements

| **Statement** | **27. The European Union should be involved in the efforts of Greece to mitigate the threat of the deepfakes online.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 3.98 |
| **Standard deviation** | 1.012 |
| **Median** | 4 |
| **Range** | 4 |
| **Total** | 123 |

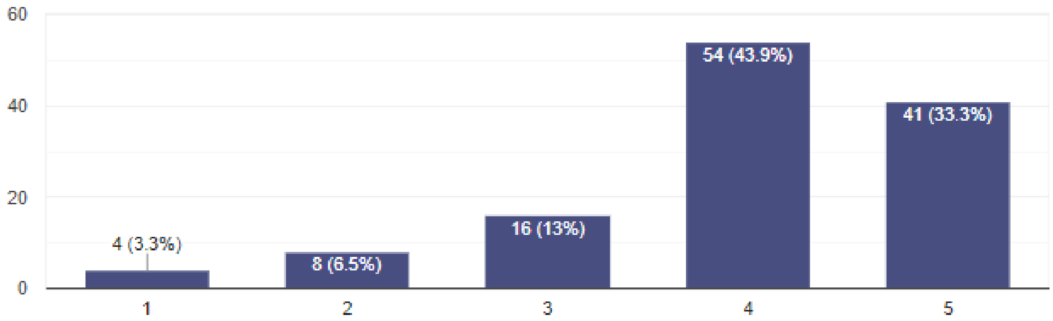


Figure C-27. Twenty-seventh statement response percentages

The overall tendency is agreement as calculated with the median of 4. This means that the users incline towards the cooperation between Greece and the EU to resolve the spread of malicious deepfakes.

The statement "Addressing deepfakes is a responsibility solely of the Greek government. The EU and non-Greek entities should not be involved" did not gather much agreement from the questionnaire responders. Specifically, as seen in the Tables C-57 and C-58 and Figure C-28, most individuals disagreed or strongly disagreed with the above statement. Collectively, 81 out of 123 responders do not think that the fight against deepfakes is a task which should be addressed not only by the Greek government but also by the EU and organisations which surpass the national level. The amount of 1 and 2 responses results in a median of 2 and general tendency of "Disagree".

Table C-57. Twenty-eight statement response frequencies

| **28. Addressing deepfakes is a responsibility solely of the Greek government. The EU and non-Greek entities should not be involved.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 37 | 44 | 28 | 13 | 1 |
| **Percentages** | 30.1% | 35.8% | 22.8% | 10.6% | 0.8% |

Table C-58. Twenty-eight statement measurements

| **Statement** | **28. Addressing deepfakes is a responsibility solely of the Greek government. The EU and non-Greek entities should not be involved.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 2.16 |
| **Standard deviation** | 1.003 |
| **Median** | 2 |
| **Range** | 4 |
| **Total** | 123 |

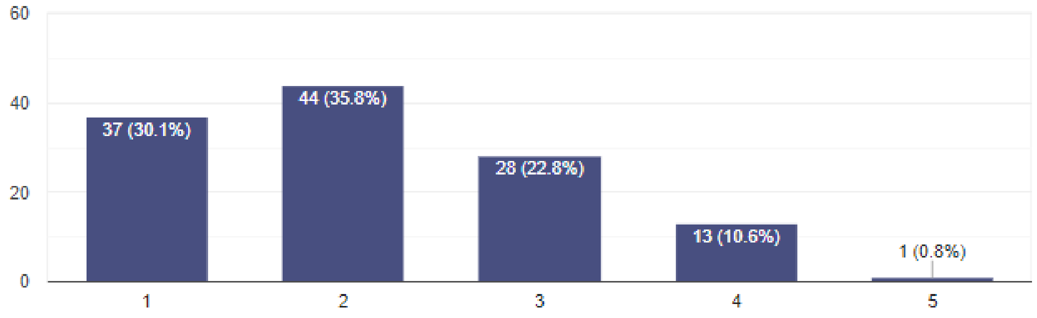


Figure C-28. Twenty-eight statement response percentages

Moving on to the statement “Greece should shape a national deepfakes strategy to address the issue of malicious deepfakes and their impact on cybersecurity”, the responders had a clear view on this matter. As seen in the Tables C-59 and C-60 and Figure C-29, most people tend to agree with the need of a national deepfakes strategy in Greece”.

Table C-59. Twenty-nineth statement response frequencies

| **29. Greece should shape a national** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **deepfakes strategy to address the issue of malicious deepfakes and their impact on cybersecurity.** | | | | | |
| **Frequencies** | 2 | 3 | 33 | 57 | 28 |
| **Percentages** | 1.6% | 2.4% | 26.8% | 46.3% | 22.8% |

Table C-60. Twenty-nineth statement measurements

| **Statement** | **29. Greece should shape a national deepfakes strategy to address the issue of malicious deepfakes and their impact on cybersecurity.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 3.86 |
| **Standard deviation** | 0.852 |
| **Median** | 4 |
| **Range** | 4 |
| **Total** | 123 |

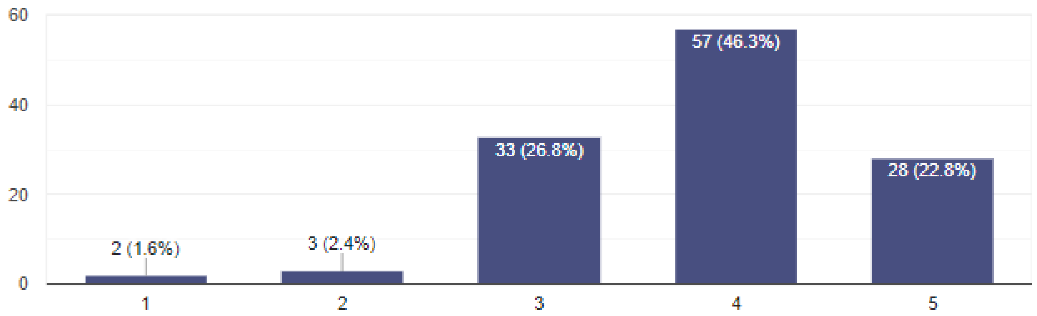


Figure C-29. Twenty-nineth statement response percentages

57 out of 123 responders agrees that Greece should shape a national deepfakes strategy. This corresponds to 46.3%. The median of 4 confirms this tendency.

The users were asked to determine the level of their agreement with the statement that the future holds more threats from deepfakes. As 44 of them agree with the statement and 45 strongly agree, this draws a clear picture of a tendency towards agreement. The following Tables C-61 and C-62 and Figure C-30 depict the frequencies of each response in more detail.

Table C-61. Thirtieth statement response frequencies

| **30. In the future, more malicious deepfakes will target users and undermine their cybersecurity.** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|
| **Frequencies** | 1 | 1 | 32 | 44 | 45 |
| **Percentages** | 0.8% | 0.8% | 26% | 35.8% | 36.6% |

Table C-62. Thirtieth statement measurements

| **Statement** | **30. In the future, more malicious deepfakes will target users and undermine their cybersecurity.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 4.07 |
| **Standard deviation** | 0.856 |
| **Median** | 4 |
| **Range** | 4 |
| **Total** | 123 |

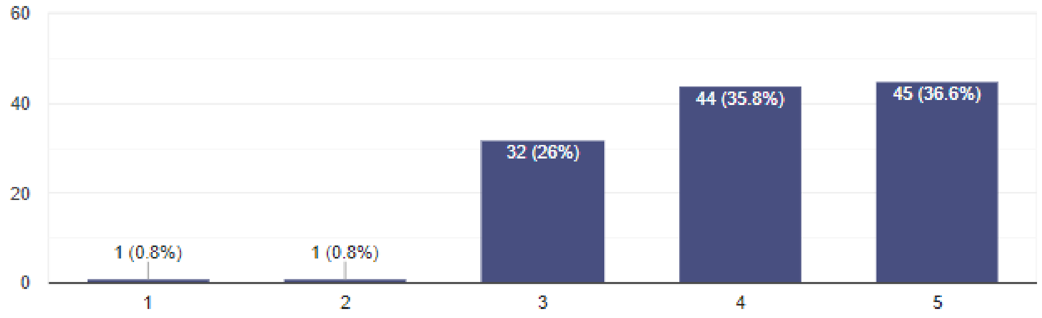


Figure C-30. Thirtieth statement response percentages

Finally, the questionnaire addressed the following statement “The threat of deepfakes is not big enough to worry now and take actions towards it”. While most users disagree with this statement, there is a strong portion of neutral responders. Specifically, 44 out of 123 participants selected the “Agree” option. 29 responders were neutral. 24 individuals strongly disagreed. This results in a median of 2, and a central inclination of disagreement with the above statement. The frequency Tables C-63 and C-64 and Figure C-31 illustrate all responses.

Table C-63. Thirty-first statement response frequencies

| **31. The threat of deepfakes is not big enough** | **Strongly disagree (1)** | **Disagree (2)** | **Neutral (3)** | **Agree (4)** | **Strongly agree (5)** |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **to worry now and take actions towards it.** | | | | | |
| **Frequencies** | 24 | 44 | 29 | 20 | 6 |
| **Percentages** | 19.5% | 35.8% | 23.6% | 16.3% | 4.9% |

Table C-64. Thirty-first statement measurements

| **Statement** | **31. The threat of deepfakes is not big enough to worry now and take actions towards it.** |
|---|---|
| **Maximum** | 5 |
| **Minimum** | 1 |
| **Mean** | 2.51 |
| **Standard deviation** | 1.126 |
| **Median** | 2 |
| **Range** | 4 |
| **Total** | 123 |

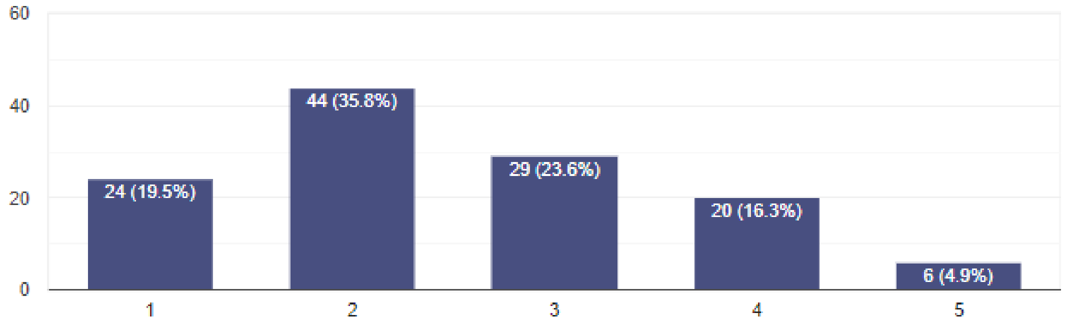


Figure C-31. Thirty-first statement response percentages

## Annex D. Deepfakes detection in Greece

As the reach of social media platforms and the Internet is global, the inherently viral nature of deepfakes can exploit this to reach a worldwide audience. Malicious actors can use deepfakes to reach and influence public opinions, spread misinformation, manipulate individuals and organisations. The threat posed by malicious deepfakes has been expressed by Greek cybersecurity experts and concurrently social media users are concerned about malicious deepfakes and feel manipulated and disoriented. The growing commodification of software for deepfakes creation can only enhance these fears as it lowers the barrier for more actors to create and disseminate malicious deepfakes. This has resulted in the detection of deepfakes becoming crucial to protect individuals and organisations in Greece from the harmful applications of deepfakes. A combination of deepfakes detection technology and human detection abilities is applied in Greece.

In this Annex, two methods of deepfakes detection in Greece are discussed. The first method relies on technology and algorithms for deepfakes detection and has been developed and applied by Facebook. The second method leans on human detection capabilities to detect fake content posted on Greek social media and websites. This detection approach is followed by Ellinika Hoaxes.

### Facebook - Michigan State University Artificial Intelligence algorithm FB-MSU AI tool

In 2021, Facebook launched a partnership with Michigan State University with the aim to develop and apply a deepfakes detection method based on reverse image engineering and AI. The method incorporates two steps a) image attribution and b) AI model discovery [162]. The procedure does not just identify the content as deepfake, but it goes a step beyond that. It uncovers the AI model which was used to generate the deepfake. This means that it may discover deepfake generation algorithms that have not been encountered before during training. This method can detect deepfakes online but also investigate instances and illegal sources of coordinated disinformation campaigns or other malicious attacks launched using deepfakes [162].

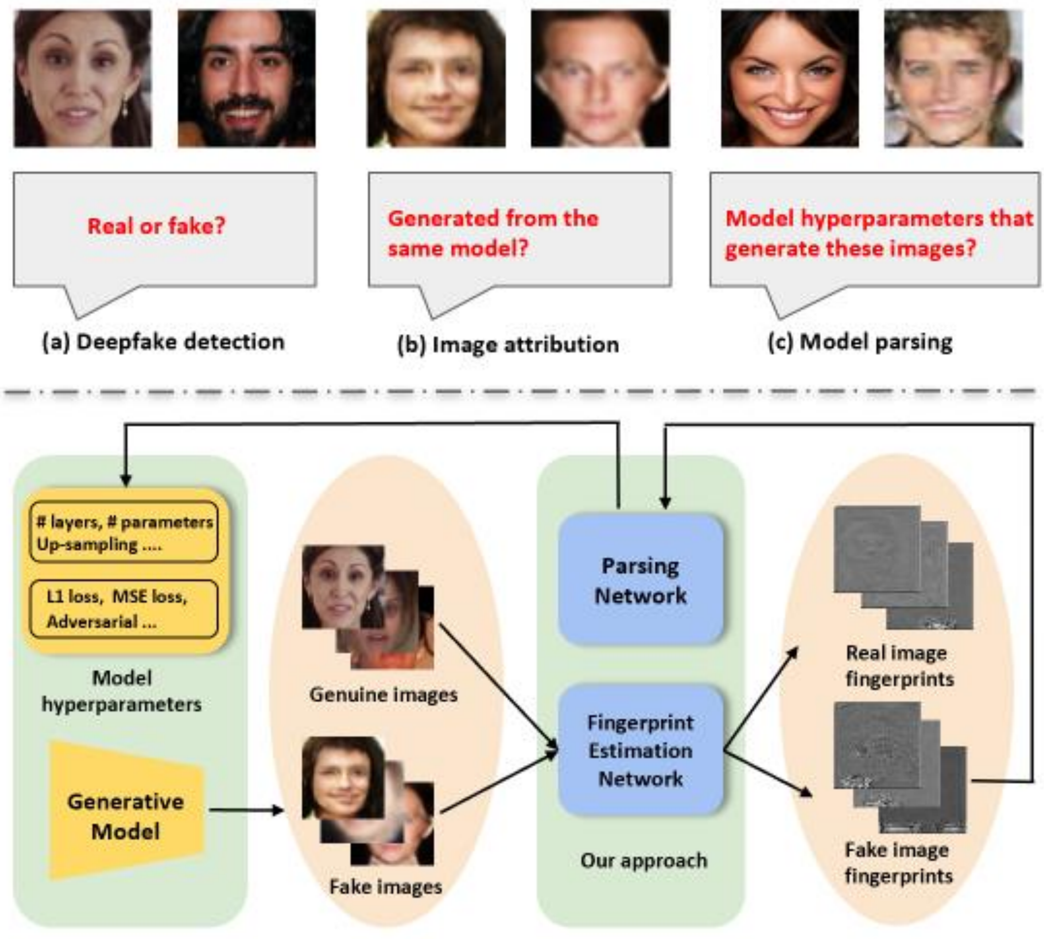


Figure D-1. Facebook - Michigan State University deepfakes detection method *[163]*

Figure D-1, as above, was retrieved from the official GitHub account of the researchers and shows the method developed by Facebook. The methodology does not require the algorithm to determine if a source content is deepfake or authentic image. Instead, the algorithm's task is to reverse engineer fingerprints, key elements and unique patterns of the AI software used to generate the content. According to Xin and Hassner, "fingerprints are unique patterns left on content generated by a generative model that can equally be used to identify the generative model that the image came from" [162]. These fingerprints are used as inputs for model parsing to predict a model's components, hyperparameters, and architecture such as the number of layers in the neural network.

Digital photography and forensic science are already using this approach to discover the device fingerprints from a single image [164]. In a similar manner, researchers from Facebook determine the generative model from the deepfake fingerprint. Yet, this approach is unique and novel in terms of deepfakes detection. According to the experts from Facebook and Michigan State University, the proposed tool scored 70% of accuracy a key benchmark test [165]. At the same time, they highlight that it is more successful than any previous tool they tested.

### Ellinika Hoaxes

Ellinika Hoaxes is a fact-checking media outlet in Greece. Its goal is to curb the spread of propaganda and disinformation in Greek media, social networking platforms, and Internet. They fact-check a wide range of content, such as politics, public health, migration, and e-commerce. Manipulated videos and images are also within the scope of their moderating activities [166]. It constitutes the only Greek organisation which fights against misinformation and the only coordinated effort to identify untrue and/or fabricated online content [166].

The methodology relies on public data, reviewing news and published articles authenticating videos and images with the use of Google images, maps and street view and image reverse search. Reading scientific studies, consulting experts and scientists are also included in the methodology [166].

The organisation is a member of the International Fact-Checking Network [166]. This guarantees the group's political independence, transparency of their funding, methodology, and its commitment to open and truthful content checking [166].  From 2019, the organisation serves as a fact-checking partners for Facebook within the fact-checking program of the social media giant [108]. Their responsibilities are to identify, review and act on false content.

According to the results, which are regularly posted on the official website of Ellinika Hoaxes, their members exposed 24 images and 7 videos, circulating on Greek internet and social media, as deepfakes [166].

The members of Ellinika Hoaxes were approached and asked to be tested upon their deepfakes detection abilities. The 7 members of the organisation were shown the same 10 images and 3 videos as the survey responders. The results can be viewed in the following Figure D-2.

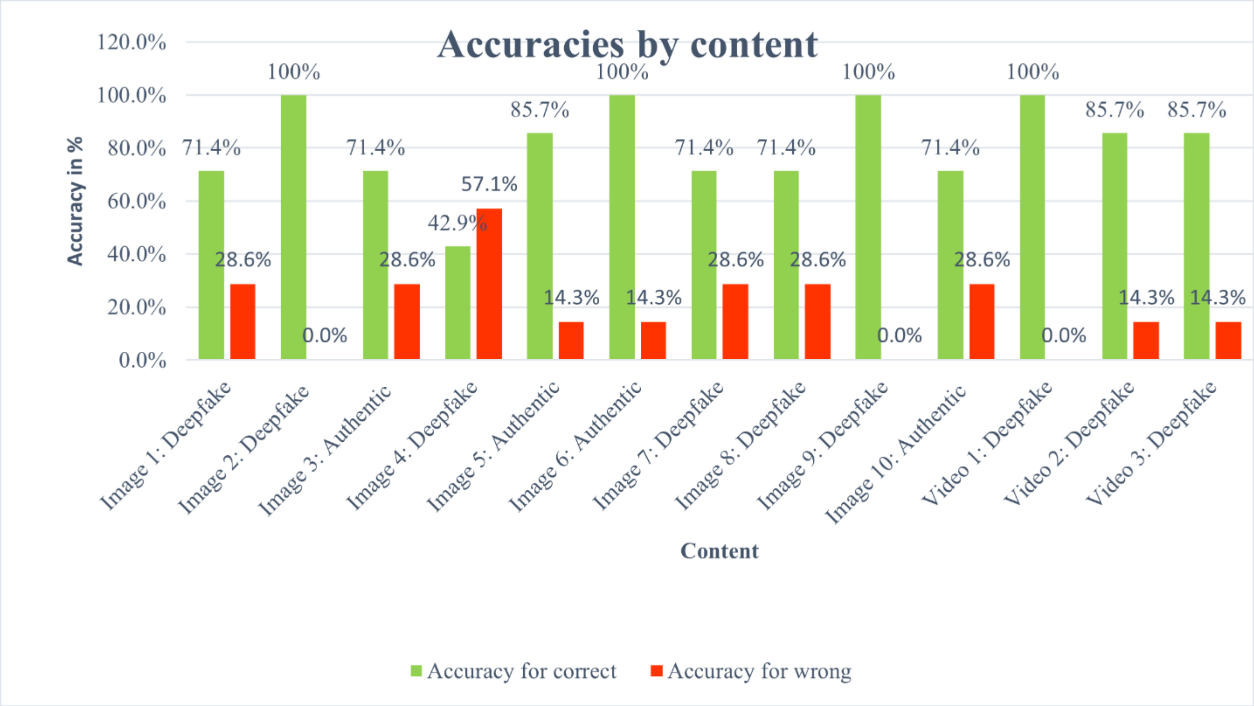


Figure D-2. Ellinika Hoaxes detection accuracies by content

Images 2, 6, 9 and Video 1 were correctly detected by all Ellinika Hoaxes participants. The lowest detection accuracy is registered in the case of Image 4. This result is akin to the result scored by individuals who are not members of Ellinika Hoaxes. They scored 31.7% detection accuracy (Figure 15). Overall, the detection accuracy of the 7 members of Ellinika Hoaxes was higher. The following Table D-1 shows a comparison of the accuracies of the two responder groups per content. The Table D-2 depicts the average detection accuracies for both groups.

Table D-1. Comparison of accuracies per content and group

| Content | Other individuals | Ellinika Hoaxes members |
|---|---|---|
| **Image 1: Deepfake** | 39,8% | 71.4% |
| **Image 2: Deepfake** | 85,4% | 100% |
| **Image 3: Authentic** | 74,8% | 71.4% |
| **Image 4: Deepfake** | 31,7% | 42.9% |
| **Image 5: Authentic** | 86,2% | 85.7% |

| | | |
|---|---|---|
| **Image 6: Authentic** | 81,1% | 100% |
| **Image 7: Deepfake** | 74,8% | 71.4% |
| **Image 8: Deepfake** | 35,0% | 71.4% |
| **Image 9: Deepfake** | 91,9% | 100% |
| **Image 10: Authentic** | 39,8% | 71.4% |
| **Video 1: Deepfake** | 91,9% | 100% |
| **Video 2: Deepfake** | 81,3% | 85.7% |
| **Video 3: Authentic** | 87,8% | 85.7% |

Table D-2. Comparison of average detection accuracies per group

| **Content** | **Average detection accuracy of all others** | **Average detection accuracy of Ellinika Hoaxes members** |
|---|---|---|
| **Images** | 64,1% | 78.6% |
| **Videos** | 87,0% | 90,5% |
| **Images + Videos** | **69,3%** | **81,3%** |

The accuracies, per content and on average, of Ellinika Hoaxes members are significantly higher than the accuracies scored by individuals who are not part of the organisation. This could indicate that their experience, methodology used at their work and continuous contact with manipulated online content enables make them more successful in discovering the signs of deepfakes. Since filtering content is their daily job routine, they are more knowledgeable and aware of typical signs of deepfakes, as well as different methods and sources to cross-check the information and images.

# References

[1] H. Pieterse and J. Botha, "Fake News and Deepfakes: A Dangerous Threat for 21st Century Information Security," March 2020. [Online]. Available: https://www.researchgate.net/publication/341454354_Fake_News_and_Deepfakes_A_Dangerous_Threat_for_21st_Century_Information_Security. [Accessed 10 October 2021].

[2] J. Donegan, "Deepfakes Are on the Rise, but Don't Panic Just Yet," Dark Reading, 6 October 2021. [Online]. Available: https://www.darkreading.com/application-security/deepfakes-are-on-the-rise-but-dont-panic-just-yet/a/d-id/1341205. [Accessed 25 October 2021].

[3] O. F. Civieta and J. Ravindran, "FBI warns of the rise of 'deepfakes' in coming months and explains how to spot them easily," 29 March 2021. [Online]. Available: https://www.businessinsider.com/fbi-investigation-generated-computer-ai-artificial-intelligence-abuse-misinformation-porn-2021-3?IR=T. [Accessed 10 October 2021].

[4] J. Kite-Powell, "The Rise Of Voice Cloning And DeepFakes In The Disinformation Wars," Forbes, 21 September 2021. [Online]. Available: https://www.forbes.com/sites/jenniferhicks/2021/09/21/the-rise-of-voice-cloning-and-deep-fakes-in-the-disinformation-wars/. [Accessed 10 October 2021].

[5] D. Castro, "Deepfakes Are on the Rise — How Should Government Respond?," Government Technology, January/February 2020. [Online]. Available: https://www.govtech.com/policy/deepfakes-are-on-the-rise-how-should-government-respond.html. [Accessed 10 October 2021].

[6] R. Toews, "Deepfakes Are Going To Wreak Havoc On Society. We Are Not Prepared.," Forbes, 25 May 2021. [Online]. Available: https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/. [Accessed 10 October 2021].

[7] Reuters, "China seeks to root out fake news and deepfakes with new online content rules," 29 November 2019. [Online]. Available: https://www.reuters.com/article/us-china-technology-idUSKBN1Y30VU. [Accessed 25 October 2021].

[8] K. Artz, "Texas Outlaws 'Deepfakes'—but the Legal System May Not Be Able to Stop Them," Law.com, 11 October 2019. [Online]. Available: https://www.law.com/texaslawyer/2019/10/11/texas-outlaws-deepfakes-but-the-legal-system-may-not-be-able-to-stop-them/?slreturn=20210925093220. [Accessed 25 October 2021].

[9] A. Jaiman, "Positive Use Cases of Deepfakes," August 2019. [Online]. Available: https://towardsdatascience.com/positive-use-cases-of-deepfakes-49f510056387. [Accessed 18 August 2021].

[10] "The Business of Fraud: Deepfakes, Fraud's Next Frontier," Recorded Future, 29 April 2021. [Online]. Available: https://go.recordedfuture.com/hubfs/reports/cta-2021-0429.pdf. [Accessed 31 October 2021].

[11] T. Brewster, "Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find," Forbes, 14 October October. [Online]. Available: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=266f4bd27559. [Accessed 31 October 2021].

[12] C. Stupp, "Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case," The Wall Street Journal, 30 August 2019. [Online]. Available: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 . [Accessed 31 October 2021].

[13] K. Quach, "Politically linked deepfake LinkedIn profile sparks spy fears, Apple cooks up AI transfer tech, and more," 17 June 2019. [Online]. Available: https://www.theregister.com/2019/06/17/roundup_ai/ . [Accessed 10 September 2021].

[14] H. Ajder, G. Patrini, F. Cavalli and L. Cullen, "The State of Deepfakes: Landscape, Threats, and Impact," September 2019. [Online]. Available: https://enough.org/objects/Deeptrace-the-State-of-Deepfakes-2019.pdf. [Accessed 10 September 2021].

[15] Mega TV, "MEGA Stories: Τα deep fake βίντεο και η Ελλάδα," 13 January 2021. [Online]. Available: https://www.megatv.com/2021/01/13/mega-stories-ta-deep-fake-vinteo-kai-i-ellada/. [Accessed 8 August 2021].

[16] e-Governance Academy Foundation, "Greece," n.d. [Online]. Available: https://ncsi.ega.ee/country/gr. [Accessed 19 August 2021].

[17] G. Drivas, "CYBERSECURITY IN GREECE: Facts, Current Needs & Future Perspectives," December 2020. [Online]. Available: https://www.sev.org.gr/Uploads/Documents/53423/Cybersecurity_in_Greece_Drivas_SEV.pdf. [Accessed 19 August 2021].

[18] L. Bertuzzi and O. Noyan , "Commission yearns for setting the global standard on artificial intelligence," Euractiv, 15 September 2021. [Online]. Available: https://www.euractiv.com/section/digital/news/commission-yearns-for-setting-the-global-standard-on-artificial-intelligence/. [Accessed 8 November 2021].

[19] C. Tsekeris, N. Demertzis, A. Linardis, O. Papaliou, A. Frangiskou, D. Kondyli and K. Iliou, "Investigating the Internet in Greece: Findings from the World Internet Project," November 2020. [Online]. Available: https://www.researchgate.net/publication/346509128_Investigating_the_Internet_in_Greece_Findings_from_the_World_Internet_Project. [Accessed 20 August 2021].

[20] A. Engler, "Fighting deepfakes when detection fails," Bookings, 14 November 2019. [Online]. Available: https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/. [Accessed 8 November 2021].

[21] J. Vincent, "Deepfake detection algorithms will never be enough," The Verge, 27 June 2019. [Online]. Available: https://www.theverge.com/2019/6/27/18715235/deepfake-detection-ai-algorithms-accuracy-will-they-ever-work. [Accessed 8 November 2021].

[22] M. Minevich, "How To Combat The Dark Side Of AI," Forbes, 28 February 2021. [Online]. Available: https://www.forbes.com/sites/markminevich/2020/02/28/how-to-combat-the-dark-side-of-ai/. [Accessed 16 November 2021].

[23] H. Ajder, G. Patrini, F. Cavalli and . L. Cullen, "The State of Deepfakes: Landscape, Threats, and Impact," September 2019. [Online]. Available: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf. [Accessed 20 August 2021].

[24] D. Karantzeni, "The political character of Social Media: How do Greek Internet users perceive and use social networks?," 30 August 2014. [Online]. Available: https://foreignpolicynews.org/2014/08/30/political-character-social-media-greek-internet-users-perceive-use-social-networks/. [Accessed 10 October 2021].

[25] Z. -. C. Belenioti, “A snapshot of Greek Social Media Users,” May 2015. [Online]. Available: https://www.researchgate.net/publication/282327907_A_snapshot_of_Greek_Social_Media_Users. [Accessed 10 October 2021].

[26] E. Lamprou, N. Antonopoulos, I. Anomeritou and C. Apostolou, "Characteristics of Fake News and Misinformation in Greece: The Rise of New Crowdsourcing-Based Journalistic Fact-Checking Models," 13 July 2021. [Online]. Available: https://doi.org/10.3390/journalmedia2030025. [Accessed 10 October 2021].

[27] D. Karantzeni, "The Political Character of Social Media: Trends, Perceptions and Prospects of Social Networking in Greece," *Journal of Self-Governance and Management Economics,* vol. 3, no. 1, p. 45–59, 2015.

[28] R. Chesney and D. K. Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," December 2019. [Online]. Available: https://scholarship.law.bu.edu/faculty_scholarship/640/ . [Accessed 2 September 2021].

[29] The Conversation, "3.2 billion images and 720,000 hours of video are shared online daily. Can you sort real from fake?," n.d. [Online]. Available: https://theconversation.com/3-2-billion-images-and-720-000-hours-of-video-are-shared-online-daily-can-you-sort-real-from-fake-148630. [Accessed 2 September 2021].

[30] DOMO, Artist, *Data Never Sleeps 5.0.* [Art]. n.d.

[31] J. Bagadiya, "367 Social Media Statistics You Must Know In 2021," 2021. [Online]. Available: https://www.socialpilot.co/blog/social-media-statistics#fb-usage-stats. [Accessed 2 September 2021].

[32] M. Albright and S. McCloskey, "Detecting GAN-generated Imagery using Color Cues," 18 December 2018. [Online]. Available: https://arxiv.org/pdf/1812.08247.pdf . [Accessed 4 September 2021].

[33] J. Lin, "You'll Love This: I'm A Beauty Writer And Tatcha's Cleansing Oil Is One Of My Skincare Must-Haves (Psst, It's 20% Off Right Now)," 17 September 2021. [Online]. Available: https://www.forbes.com/sites/forbes-personal-shopper/2021/09/17/tatcha-cleansing-oil-review/?sh=5dadd18632af. [Accessed 4 September 2021].

[34] N. Martin, "How Social Media Has Changed How We Consume News," 30 November 2018. [Online]. Available: https://www.forbes.com/sites/nicolemartin1/2018/11/30/how-social-media-has-changed-how-we-consume-news/?sh=692091bf3c3c . [Accessed 4 September 2021].

[35] N. Newman, R. Fletcher, A. Schulz, S. Andı and R. K. Nielsen, "Reuters Institute Digital News Report 2020," n.d. [Online]. Available:

https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf. [Accessed 4 September 2021].

[36] E. Shearer and A. Mitchell, "News Use Across Social Media Platforms in 2020," 12 January 2021. [Online]. Available: https://www.pewresearch.org/journalism/2021/01/12/news-use-across-social-media-platforms-in-2020/ . [Accessed 4 September 2021].

[37] M. Vorhaus, "People Increasingly Turn To Social Media For News," 24 June 2020. [Online]. Available: https://www.forbes.com/sites/mikevorhaus/2020/06/24/people-increasingly-turn-to-social-media-for-news/ . [Accessed 4 September 2021].

[38] Y. Li, M.-C. Chang and S. Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking," 11 June 2018. [Online]. Available: https://arxiv.org/pdf/1806.02877.pdf. [Accessed 5 September 2021].

[39] S. Pattanayak, "Deepfake: Is it The New Weapon Of Choice?," 4 October 2020. [Online]. Available: https://odishabytes.com/deepfake-is-it-the-new-weapon-of-choice/. [Accessed 4 September 2021].

[40] B. Wittes and G. Blum, The future of violence: robots and germs, hackers and drones—confronting a new age of threat, New York: Basic Books,, 2015.

[41] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," November 2018. [Online]. Available: https://engineering.purdue.edu/~dgueraco/content/deepfake.pdf. [Accessed 4 September 2021].

[42] A. Jaiman, "Positive Use Cases of Deepfakes," 14 August 2020. [Online]. Available: https://towardsdatascience.com/positive-use-cases-of-deepfakes-49f510056387 . [Accessed 4 September 2021].

[43] L. Cerulus, “One group that’s embraced AI: Criminals,” 30 May 2021 . [Online]. Available: https://www.politico.eu/article/artificial-intelligence-criminals/ . [Accessed 4 September 2021].

[44] M. Hogan, “REPLICATING REALITY Advantages and Limitations of Weaponized Deepfake Technology,” April 2020. [Online]. Available: https://www.wm.edu/offices/global-research/research-labs/pips/white_papers/2019-2020/hogan-final.pdf. [Accessed 4 September 2021].

[45] E. Thompson , Artist, *What is fire triangle?.* [Art]. Gustavb, National Park Service , NOAA and University of Washington, n.d .

[46] T. T. Nguyen, Q. V. Hung Nguyen, C. M. Nguyen, D. Nguyen, D. T. Nguyen and S. Nahavand, “Deep Learning for Deepfakes Creation and Detection: A Survey,” 26 April 2021. [Online]. Available: https://arxiv.org/pdf/1909.11573.pdf. [Accessed 5 September 2021].

[47] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Moralez and J. Ortega-Garcia, “Deepfakes and beyond: A Survey of face manipulation and fake detection,” December 2020. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1566253520303110. [Accessed 4 September 2021].

[48] T. Hwang, “The Deepfakes: A Grounded Threat Assessment,” 2020. [Online]. Available: https://cset.georgetown.edu/wp-content/uploads/CSET-Deepfakes-Report.pdf. [Accessed 4 September 2021].

[49] J. Kietzmann, T. C. Kietzmann and L. W. Lee, "Deepfakes: Trick or Treat?," n.d. [Online]. Available: https://core.ac.uk/download/pdf/250590695.pdf. [Accessed 4 September 2021].

[50] Belfer Center for Science and International Affairs, "TECH POLICY FACTSHEET: DEEPFAKES," 2020. [Online]. Available: https://www.belfercenter.org/sites/default/files/2020-10/tappfactsheets/Deepfakes.pdf. [Accessed 4 September 2021].

[51] R. Metz, "The fight to stay ahead of deepfake videos before the 2020 US election," 12 June 2019. [Online]. Available:

https://edition.cnn.com/2019/06/12/tech/deepfake-2020-detection/index.html. [Accessed 5 September 2021].

[52] G. Shaor, "What ‘deepfakes’ are and how they may be dangerous," 17 January 2020. [Online]. Available: https://www.cnbc.com/2019/10/14/what-is-deepfake-and-how-it-might-be-dangerous.html. [Accessed 5 September 2021].

[53] A. G. Johansen, “Deepfakes: What they are and why they’re threatening,” 24 July 2020. [Online]. Available: https://us.norton.com/internetsecurity-emerging-threats-what-are-deepfakes.html. [Accessed 5 September 2021].

[54] I. Sample, "What are deepfakes – and how can you spot them?," 13 January 2020. [Online]. Available: https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them. [Accessed 5 September 2021].

[55] R. Davis, "Technology Factsheet: Deepfakes," 2020. [Online]. Available: https://www.belfercenter.org/publication/technology-factsheet-deepfakes#:~:text=Deepfakes%20can%20be%20defined%20as%20synthetic%20auditory%20or,often%20created%20with%20the%20intent%20of%20deceiving%20audiences.. [Accessed 5 September 2021].

[56] J. Donovan and B. Paris, "Deep Fakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence," n.d. [Online]. Available: https://datasociety.net/wp-content/uploads/2019/09/DataSociety_Deepfakes_Cheap_Fakes.pdf. [Accessed 7 September 2021].

[57] Cannyai, "Transform Your Videos to any Language, with any Dialogue," n.d. [Online]. Available: https://www.cannyai.com/ . [Accessed 8 October 2021].

[58] Adobe, “Adobe After Effects,” n.d. [Online]. Available: https://www.adobe.com/products/aftereffects.html. [Accessed 7 September 2021].

[59] Vegas, "VEGAS Pro," n.d. [Online]. Available: https://www.vegascreativesoftware.com/us/vegas-pro/. [Accessed 7 September 2021].

[60] M. Westerlund, “The Emergence of Deepfake Technology: A Review,” November 2019. [Online]. Available: https://timreview.ca/article/1282 . [Accessed 10 September 2021].

[61] Y. Nirkin, Y. Keller and T. Hassner, “FSGAN: Subject Agnostic Face Swapping and Reenactment,” n.d. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2019/papers/Nirkin_FSGAN_Subject_Agnostic_Face_Swapping_and_Reenactment_ICCV_2019_paper.pdf. [Accessed 10 September 2021].

[62] The Byte, “NEW SYSTEM MAKES IT TROUBLINGLY EASY TO CREATE DEEPFAKES FSGAN CAN SWAP FACES IN REAL-TIME, NO TRAINING REQUIRED.,” n.d. [Online]. Available: https://futurism.com/the-byte/system-easy-create-deepfakes. [Accessed 10 September 2021].

[63] Techwiser, “How to Use DeepFaceLab on Windows to Create Your First Deepfake?,” 10 September 2019. [Online]. Available: https://techwiser.com/how-to-use-deepfacelab-on-windows/. [Accessed 10 September 2021].

[64] DFBlue, "The DeepFaceLab Tutorial (always up-to-date)," 25 October 2019. [Online]. Available: https://pub.dfblue.com/pub/2019-10-25-deepfacelab-tutorial. [Accessed 10 September 2021].

[65] R. Senaratne, “Make Your Own Deepfake Video in a Few Easy Steps,” 13 July 2020 · . [Online]. Available: https://heartbeat.comet.ml/make-your-own-deepfake-video-in-a-few-easy-steps-fd15824624ac. [Accessed 10 September 2021].

[66] A. Siarohin, "First Order Motion Model for Image Animation," 2019. [Online]. Available: https://github.com/AliaksandrSiarohin/first-order-modelAliaksandrSiarohin. [Accessed 10 September 2021].

[67] Deepfakes Web, “Online Deepfake Maker,” n.d. [Online]. Available: https://deepfakesweb.com/. [Accessed 10 September 2021].

[68] Malavida, "FakeApp," n.d. [Online]. Available: https://www.malavida.com/en/soft/fakeapp/. [Accessed 10 September 2021].

[69] C. Vaccari and A. Chadwick, “Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News,” 19 February 2020. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/2056305120903408. [Accessed 10 September 2021].

[70] J. Brownlee, "How Do Convolutional Layers Work in Deep Learning Neural Networks?," 17 April 2019. [Online]. Available: https://machinelearningmastery.com/convolutional-layers-for-deep-learning-neural-networks/ . [Accessed 10 September 2021].

[71] I. Laptev, B. Rozenfeld, C. Schmid and . M. Marszałek, “Learning Realistic Human Actions from Movies,” 2008. [Online]. Available: https://www.di.ens.fr/~laptev/download.html#actionclassification. [Accessed 10 October 2021].

[72] O. Giudice, L. Guarnera and S. Battiato, "Fighting Deepfakes by Detecting GAN DCT Anomalies," January 2021. [Online]. Available: https://arxiv.org/pdf/2101.09781.pdf. [Accessed 10 October 2021].

[73] L. Guarnera, O. Giudice and S. Battiato, "DeepFake Detection by Analyzing Convolutional Traces," 2020. [Online]. Available: https://arxiv.org/pdf/2004.10448.pdf. [Accessed 9 October 2021].

[74] U. A. Ciftci, I. Demir and L. Yin, "How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological

Signals," August 2021. [Online]. Available: https://arxiv.org/pdf/2008.11363.pdf. [Accessed 8 October 2021].

[75] Analystics India Magazine, "Machine Learning Researchers Spot Deep Fakes From Heartbeats," 10 September 2020. [Online]. Available: https://analyticsindiamag.com/deepfake-heartbeat-machine-learning-detection/. [Accessed 8 October 2021].

[76] S. Agarwal, O. Fried and H. Farid, "Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches," 2020. [Online]. Available: https://www.ohadf.com/papers/AgarwalFaridFriedAgrawala_CVPRW2020.pdf. [Accessed 8 October 2021].

[77] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano and H. Li, "Protecting World Leaders Against Deep Fakes," 2020. [Online]. Available: http://www.hao-li.com/publications/papers/cvpr2019workshopsPWLADF.pdf. [Accessed 8 October 2021].

[78] A. Pishori, B. Rollins, N. van Houten, N. Chatwani and O. Uraimov, "Detecting Deepfake Videos: An Analysis of Three Techniques," 20 July 2020. [Online]. Available: https://arxiv.org/pdf/2007.08517.pdf. [Accessed 7 October 2021].

[79] S. Scott McCloskey and M. Albright, "Detecting GAN-generated Imagery using Color Cues," 19 December 2018. [Online]. Available: https://arxiv.org/pdf/1812.08247.pdf. [Accessed 7 October 2021].

[80] M. Groh, Z. Epstein, N. Obradovich, M. Cebrian and I. Rahwan, "Human detection of machine manipulated media," 8 November 2019. [Online]. Available: https://arxiv.org/pdf/1907.05276.pdf. [Accessed 7 October 2021].

[81] N. M. Muller, K. Markert and K. Bottinger, "Human Perception of Audio Deepfakes," 30 September 2021. [Online]. Available: https://arxiv.org/pdf/2107.09667.pdf. [Accessed 7 October 2021].

[82] P. Korshunov and S. Marcel, "Deepfake detection: humans vs. machines," 7 September 2020. [Online]. Available: https://arxiv.org/pdf/2009.03155.pdf . [Accessed 7 October 2021].

[83] M. Groh, Z. Epstein, C. Firestone and . R. Picard, "Comparing Human and Machine Deepfake Detection with Affective and Holistic Processing," 13 May 2021. [Online]. Available: https://arxiv.org/pdf/2105.06496.pdf. [Accessed 8 October 2021].

[84] K. Canales, "Facebook's largest content moderator has reportedly struggled with the ethics of its work for the company, which requires contractors to sift through violent, graphic content," 31 August 2021. [Online]. Available: https://www.businessinsider.com/facebook-content-moderation-accenture-questioned-ethics-2021-8?IR=T. [Accessed 8 October 2021].

[85] Y. Papadopoulos, "Inside Facebook's moderation hub in Athens," 9 November 2019. [Online]. Available: https://www.ekathimerini.com/in-depth/special-report/246279/inside-facebook-s-moderation-hub-in-athens/. [Accessed 19 August 2021].

[86] M. Bickert, "Enforcing Against Manipulated Media," 6 January 6, 2020 2020. [Online]. Available: https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/. [Accessed 19 August 2021].

[87] Ellinika Hoaxes, "Ερωτήσεις σχετικές με το Facebook," n.d. [Online]. Available: https://www.ellinikahoaxes.gr/facebook-and-ellinika-hoaxes/. [Accessed 19 August 2021].

[88] "Athens At The Center Of European Cyber Security Strategy," 10 February 2019. [Online]. Available: https://www.forbes.com/sites/yiannismouratidis/2019/02/10/athens-at-the-center-of-european-cyber-security-strategy/?sh=7050d7df239b. [Accessed 19 August 2021].

[89] e-Governance Academy Foundation, "Methodology," n.d. [Online]. Available: https://ncsi.ega.ee/methodology/. [Accessed 19 August 2021].

[90] EBU Media Intelligence Service, "Trust in Media 2020," April 2020. [Online]. Available: https://medienorge.uib.no/files/Eksterne_pub/EBU-MIS-Trust_in_Media_2020.pdf. [Accessed 19 August 2021].

[91] EUR-Lex , n.d. [Online]. Available: https://eur-lex.europa.eu/homepage.html. [Accessed 10 October 2021].

[92] European Union, "Official website of the European Union," n.d. [Online]. Available: https://europa.eu/european-union/index_en . [Accessed 10 October 2021].

[93] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci and N. Sebe, "First Order Motion Model for Image Animation," 2019. [Online]. Available: https://aliaksandrsiarohin.github.io/first-order-model-website/. [Accessed 8 October 2021].

[94] A. Siarohin, "Demo for paper "First Order Motion Model for Image Animation"," 2019. [Online]. Available: https://colab.research.google.com/github/AliaksandrSiarohin/first-order-model/blob/master/demo.ipynb#scrollTo=Oxi6-riLOgnm. [Accessed 8 October 2021].

[95] FaceApp , "Most Popular Selfie Editor," n.d. [Online]. Available: https://www.faceapp.com/ . [Accessed 8 October 2021].

[96] D. Thomas, "Deepfakes: A threat to democracy or just a bit of fun?," BBC, 23 January 2020. [Online]. Available: https://www.bbc.com/news/business-51204954. [Accessed 10 October 2021].

[97] I. Perov, "DeepFaceLab," 2019. [Online]. Available: https://github.com/iperov/DeepFaceLab. [Accessed 8 October 2021].

[98] Awesome Open Source, "Awesome Open Source," n.d. [Online]. Available: https://awesomeopensource.com/project/iperov/DeepFaceLab . [Accessed 8 October 2021].

[99] Shamook, Director, *Tom Cruise is Iron Man [DeepFake].* [Film]. 2020.

[100] *Robert Downey Jr and Tom Holland in Back to the future - This is heavy! [ deepfake ].* [Film]. 2020.

[101] M. Tavakol and R. Dennick, "Making sense of Cronbach's alpha," 27 June 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4205511/ . [Accessed 10 October 2021].

[102] A. Joshi, S. Kale, S. Chandel and D. K. Pal, "Likert Scale: Explored and Explained," *Current Journal of Applied Science and Technology,* vol. 7, no. 4, pp. 396-403, 2015.

[103] D. Madrigal and B. McClain, "Strengths and Weaknesses of Quantitative and Qualitative Research," 3 September 2021. [Online]. Available: https://www.uxmatters.com/mt/archives/2012/09/strengths-and-weaknesses-of-quantitative-and-qualitative-research.php . [Accessed 10 October 2021].

[104] F. Almeida, D. Faria and A. Queirós, "Strengths and Limitations of Qualitative and Quantitative Research Methods," September 2017. [Online]. Available: https://www.researchgate.net/publication/319852576_Strengths_and_Limitations_of_Qualitative_and_Quantitative_Research_Methods. [Accessed 10 October 2021].

[105] V. O. Ajayi, "Primary Sources of Data and Secondary Sources of Data," September 2017. [Online]. Available: https://www.researchgate.net/publication/320010397_Primary_Sources_of_Data_and_Secondary_Sources_of_Data . [Accessed 10 October 2021].

[106] S. M. S. Kabir, "METHODS OF DATA COLLECTION," in *Basic Guidelines for Research: An Introductory Approach for All Disciplines*, Chittagong, Book Zone Publication, 2016, p. Chapter 9.

[107] Α. Λακασάς, "«Ανοχύρωτη» στα fake news η Ελλάδα," 28 July 2021. [Online]. Available: https://www.kathimerini.gr/society/561448705/anochyroti-sta-fake-news-i-ellada/. [Accessed 22 August 2021].

[108] Ellinika Hoaxes, "Facebook announces its fact checker partner in Greece," May 2019. [Online]. Available: https://media.ellinikahoaxes.gr/uploads/2019/05/Press-Release_Facebook-announces-its-fact-checker-partner-in-Greece.pdf. [Accessed 19 August 2021].

[109] Μ. Αθανασίου, "Το νέο όπλο κατά των fake news σε Ελλάδα, Κύπρο," 26 May 2021. [Online]. Available: https://www.kathimerini.gr/society/561376480/to-neo-oplo-kata-ton-fake-news-se-ellada-kypro/. [Accessed 19 August 2021].

[110] MSU Today, "MSU, Facebook develop research model to fight deepfakes," 16 June 2021. [Online]. Available: https://msutoday.msu.edu/news/2021/deepfake-detection . [Accessed 10 October 2021].

[111] X. Yin and T. Hassner, "Reverse engineering generative models from a single deepfake image," 16 June 2021. [Online]. Available: https://ai.facebook.com/blog/reverse-engineering-generative-model-from-a-single-deepfake-image. [Accessed 10 October 2021].

[112] X. Yin, V. Asnani, T. Hassner and X. Liu, "Reverse Engineering of Generative Models: Inferring Model Hyperparameters from Generated Images," 2021. [Online]. Available: https://github.com/vishal3477/Reverse_Engineering_GMs?fbclid=IwAR2Dx7_6R5Q8u1YKPqOPB3lEPZuFCC1H0QMM19RJNmMWliNngJR_JGB56ho. [Accessed 10 October 2021].

[113] M. Groh, Z. Epstein, C. Firestone and R. Picard, "Deepfake Detection by Human Crowds, Machines, and Machine-informed Crowds," Massachusetts

Institute of Technology Media Lab, Johns Hopkins University, 27 October 2021. [Online]. Available: https://mattgroh.com/pdfs/detect-fakes-comparing-human-machine.pdf. [Accessed 3 December 2021].

[114] S. Papathanassopoulos, "Greece Media: Social Networks," Media Landscape , n.d. [Online]. Available: https://www.medialandscapes.org/country/greece/media/social-networks. [Accessed 5 December 2021].

[115] Statista, "Market share held by the leading social networks in Greece from 2010 to 2020," March 2021. [Online]. Available: https://www.statista.com/statistics/621193/leading-social-networks-ranked-by-market-share-in-greece/#:~:text=%20Facebook%20holds%20the%20largest%20share%20of%20the,sat%20at%2053%20percent%20in%20Greece%20during%202018.. [Accessed 5 December 2021].

[116] StatCounter, "Social Media Stats Greece," 2021. [Online]. Available: https://gs.statcounter.com/social-media-stats/all/greece. [Accessed 5 December 2021].

[117] J. Foley, "14 deepfake examples that terrified and amused the internet," Creative Bloq, 1 June 2021. [Online]. Available: https://www.creativebloq.com/features/deepfake-examples. [Accessed 5 December 2021].

[118] C. Reid, "Lecture explores deepfakes, media manipulation," The Observer, 22 March 2021. [Online]. Available: https://ndsmcobserver.com/2021/03/lecture-explores-deepfakes-media-manipulation/ . [Accessed 4 December 2021].

[119] Exploring our mind, "Deepfakes, The New Form of Digital Manipulation," 15 November 2021. [Online]. Available: https://exploringyourmind.com/deepfakes-the-new-form-of-digital-manipulation/ . [Accessed December 2021].

[120] T. Kirchengast, "Deepfakes and image manipulation: criminalisation and control," 16 July 2020. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/13600834.2020.1794615?journalCode=cict20. [Accessed 4 December 2021].

[121] W. W. Daniel, "Chapter 6: Hypothesis Testing," in *Basic Concepts and Methodology for the Health Sciences*, 2009.

[122] C. Andrade, "The P Value and Statistical Significance: Misunderstandings, Explanations, Challenges, and Alternatives," *Indian J Psychol Med,* vol. 41, no. 3, pp. 210-215, May-June 2019.

[123] G. E. Dallal, "Why P=0.05?," 22 December 2020. [Online]. Available: https://www.webpages.uidaho.edu/~brian/why_significance_is_five_percent.pdf. [Accessed 22 December 2021].

[124] C. Taylor, "What 'Fail to Reject' Means in a Hypothesis Test," ThoughtCo, 28 January 28 2019. [Online]. Available: https://www.thoughtco.com/fail-to-reject-in-a-hypothesis-test-3126424. [Accessed 22 December 2021].

[125] Frank Slide, "What does P value greater than 0.05 mean?," n.d. [Online]. Available: https://www.frankslide.com/what-does-p-value-greater-than-0-05-mean/#:~:text=A%20p-value%20higher%20than%200.%2005%20%28%3E%200.,and%20indicates%20strong%20evidence%20for%20the%20null%20hypothesis.. [Accessed 22 December 2021].

[126] R. DeSantis, “Kristen Bell Recalls Shock of Learning Her Face Was Used in Pornographic Deepfake: 'It's Not OK',” People, 10 June 2020. [Online]. Available: https://people.com/human-interest/kristen-bell-shock-face-used-pornographic-deepfake/ . [Accessed 18 December 2021].

[127] S. Gonstad, "Deepfakes and the future of entertainment," Comic Years, 31 July 2019. [Online]. Available: https://comicyears.com/pop-culture/deepfakes-and-the-future-of-entertainment/ . [Accessed 18 December 2021].

[128] R. Castoro, "Deepfakes Could Mean Better Movies but a Worse Society," Collider, 7 November 2019. [Online]. Available: https://collider.com/deepfakes-explained-better-movies-worse-society/ . [Accessed 18 December 2021].

[129] CNN Greece, "Πολιτική αστάθεια και υψηλή φορολογία οι βασικές έγνοιες των επιχειρήσεων για το 2016," 28 December 2015. [Online]. Available: https://www.cnn.gr/oikonomia/epixeiriseis/story/16770/politiki-astatheia-kai-ypsili-forologia-oi-vasikes-egnoies-ton-epixeiriseon-gia-to-2016 . [Accessed 18 December 2021].

[130] O. Anastasakis, "Opinion: Has Greece's democracy regressed?," CNN, 1 March 2012. [Online]. Available: https://edition.cnn.com/2012/03/01/opinion/greece-democracy/index.html#:~:text=Greece%20is%20gradually%20losing%20its%20middle%20class%2C%20the,while%20those%20in%20the%20center%20become%20most%20vulnerable. . [Accessed 18 December 2021].

[131] Digital Skills and Jobs Platform, "Ministry of Digital Governance of Greece," n.d. [Online]. Available: https://digital-skills-jobs.europa.eu/en/organisations/ministry-digital-governance-greece. [Accessed 2 February 2022].

[132] Cyber Alert, "CYBER ALERT," n.d. [Online]. Available: https://cyberalert.gr/. [Accessed 13 January 2022].

[133] Cyber Security International Institute, "Who we are," n.d. [Online]. Available: https://www.csii.gr/poioi-eimaste/. [Accessed 13 January 2022].

[134] Eurostat, "Individuals' level of digital skills (until 2019)[isoc_sk_dskl_i]," 16 December 2021. [Online]. Available: https://appsso.eurostat.ec.europa.eu/nui/show.do?query=BOOKMARK_DS-601368_QID_-6A0D7467_UID_-3F171EB0&layout=TIME,C,X,0;GEO,L,Y,0;INDIC_IS,L,Z,0;IND_TYPE,L,Z,1;UNIT,L,Z,2;INDICATORS,C,Z,3;&zSelection=DS-

601368IND_TYPE,IND_TOTAL;DS-601368INDIC_IS,I_DSK_AB;. [Accessed 13 January 2022].

[135] CBS, "The Netherlands ranks among the EU top in digital skills," 14 February 2021. [Online]. Available: https://www.cbs.nl/en-gb/news/2020/07/the-netherlands-ranks-among-the-eu-top-in-digital-skills. [Accessed 13 January 2022].

[136] e-Estonia, "Estonia as an international cybersecurity leader," 19 August 2019. [Online]. Available: https://e-estonia.com/estonia-as-an-international-cybersecurity-leader/. [Accessed 13 January 2022].

[137] N. Seshadri and S. A, "Are Copyright Laws adequate to deal with Deepfakes?: A comparative analysis of positions in the United States, India and United Kingdom," KSLR Commercial & Financial Law Blog, 17 December 2020. [Online]. Available: https://blogs.kcl.ac.uk/kslrcommerciallawblog/2020/12/17/are-copyright-laws-adequate-to-deal-with-deepfakes-a-comparative-analysis-of-positions-in-the-united-states-india-and-united-kingdom/. [Accessed 24 November 2021].

[138] P. Nema, "Understanding copyright issues entailing deepfakes in India," *International Journal of Law and Information Technology,* vol. Volume 29, no. Issue 3, p. Pages 241–254, 2021.

[139] Hellenic Copyright Organisation , "Law 2121/1993 Copyright, Related Rights and Cultural Matters," 4 March 1993. [Online]. Available: https://www.opi.gr/index.php/en/library/law-2121-1993#a2 . [Accessed 24 November 2021].

[140] ΥΠΟΥΡΓΕΙΟ ΠΟΛΙΤΙΣΜΟΥ ΚΑΙ ΑΘΛΗΤΙΣΜΟΥ, "ΘΕΜΑ: Συγκρότηση Νομοπαρασκευαστικής Επιτροπής," 19 September 2019. [Online]. Available: https://www.opi.gr/images/library/nomothesia/ethniki/ypoyrgikes_apofaseis/12521_2019.pdf . [Accessed 24 November 2021].

[141] F. Ferri, "The dark side(s) of the EU Directive on copyright and related rights in the Digital Single Market," *China-EU Law J,* 2020.

[142] Capital, "Πιερρακάκης: Η Ελλάδα είναι έτοιμη να παρουσιάσει το Εθνικό Σχέδιο για την Τεχνητή Νοημοσύνη," 29 September 2021. [Online]. Available: https://www.capital.gr/epikairotita/3585110/pierrakakis-i-ellada-einai-etoimi-na-parousiasei-to-ethniko-sxedio-gia-tin-texniti-noimosuni . [Accessed 24 November 2021].

[143] T. Kokkinidis, "Greece Approves Bill for Tougher Penalties on Crime, Fake News," Greekreporter, 12 November 2021. [Online]. Available: https://greekreporter.com/2021/11/12/greece-tougher-penalties-crime-fake-news/ . [Accessed 24 November 2021].

[144] M. Tsimitakis and S. Michalopoulos, "Greek media landscape raises eyebrows in Brussels," EURACTIV.com , 12 November 2021. [Online]. Available: https://www.euractiv.com/section/politics/short_news/greek-media-landscape-raises-eyebrows-in-brussels/ . [Accessed 24 November 2021].

[145] N. Nikolinakos, D. Kouvelou and A. Spyropoulos , "Greece: Cybersecurity Laws and Regulations 2022," ICLG, 2021. [Online]. Available: https://iclg.com/practice-areas/cybersecurity-laws-and-regulations/greece . [Accessed 24 November 2021].

[146] V. Sakellarides, "COMPUTER RELATED CRIMES AND COMPUTER INTEGRITY CRIMES IN GREECE AND CASE LAW," n.d. [Online]. Available: https://sakellarides.gr/portal/internetlaw-en.php . [Accessed 24 November 2021].

[147] Lawspot.gr, "Άρθρο 348A - Ποινικός Κώδικας (Νόμος 4619/2019) - Πορνογραφία ανηλίκων," 1 July 2019. [Online]. Available: https://www.lawspot.gr/nomikes-plirofories/nomothesia/n-4619-2019/arthro-348a-poinikos-kodikas-nomos-4619-2019-pornografia . [Accessed 24 November 2021].

[148] EUROPEAN COMMISSION, "Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS," 21 April 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 . [Accessed 24 November 2021].

[149] GDPR , "Art. 4 GDPR Definitions," [Online]. Available: https://gdpr-info.eu/art-4-gdpr/ . [Accessed 24 November 2021].

[150] GDPR, "Art. 6 GDPRLawfulness of processing," [Online]. Available: https://gdpr-info.eu/art-6-gdpr/ . [Accessed 24 November 2021].

[151] GDPR, "Art. 7 GDPR Conditions for consent," [Online]. Available: https://gdpr-info.eu/art-7-gdpr/ . [Accessed 24 November 2021].

[152] GDPR, "Art. 9 GDPR Processing of special categories of personal data," [Online]. Available: https://gdpr-info.eu/art-9-gdpr/. [Accessed 24 November 2021].

[153] European IPR Helpdesk , "Fact Sheet Copyright Essentials," September 2017. [Online]. Available: https://www.ipoi.gov.ie/en/commercialise-your-ip/tools-for-business/copyright-essentials.pdf . [Accessed 24 November 2021].

[154] EUROPEAN COMMISSION, "Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC," 15 December 2020. [Online]. Available: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN. [Accessed 24 November 2021].

[155] . A. Tidey, A. Lazaro and J. Parrock, "Digital Services Act: Brussels vows to 'put order into chaos' of digital world with new tech laws," Euronews, 15

December 2020. [Online]. Available: https://www.euronews.com/2020/12/15/digital-services-act-brussels-unveils-landmark-plans-to-regulate-tech-companies. [Accessed 24 November 2021].

[156] UCLA STATISTICAL CONSULTING GROUP, "WHAT DOES CRONBACH’S ALPHA MEAN? | SPSS FAQ," n.d. [Online]. Available: https://stats.oarc.ucla.edu/spss/faq/what-does-cronbachs-alpha-mean/. [Accessed 15 December 2021].

[157] P. Chetty and S. Datt, "Reliability test in SPSS using Cronbach Alpha," Project Guru, 7 February 2015. [Online]. Available: https://www.projectguru.in/reliability-test-in-spss-using-cronbach-alpha/#:~:text=Cronbach%20Alpha%20is%20a%20reliability%20test%20conducted%20within,determine%20if%20the%20scale%20is%20reliable%20or%20not.. [Accessed 23 December 2021].

[158] P. R. Hilton, C. Brownlow, I. McMurray and B. Cozens, SPSS Explained, Routledge, 2004.

[159] A. Kostoulas, "Central tendencies and statistical significance in Likert scales," 19 November 2013. [Online]. Available: https://achilleaskostoulas.com/2013/11/19/likert-scales-central-tendency-statistical-significance/ . [Accessed 22 December 2021].

[160] S. Jamieson, "Likert scales: how to (ab)use them," *Med Educ,* vol. 39, no. 9, p. 970, 2005.

[161] G. M. Sullivan and A. R. Artino, "Analyzing and Interpreting Data From Likert-Type Scales," *J Grad Med Educ,* vol. 5, no. 4, p. 541–542, 2013.

[162] X. Yin and T. Hassner, "Reverse engineering generative models from a single deepfake image," 16 June 2021. [Online]. Available: https://ai.facebook.com/blog/reverse-engineering-generative-model-from-a-single-deepfake-image/. [Accessed 22 January 2022].

[163] V. Asnani, X. Yin, T. Hassner and X. Liu, “Reverse_Engineering_GMs,” 24 July 2021. [Online]. Available: https://github.com/vishal3477/Reverse_Engineering_GMs?fbclid=IwAR2Dx7_6R5Q8u1YKPqOPB3lEPZuFCC1H0QMM19RJNmMWliNngJR_JGB56ho.. [Accessed 22 January 2022].

[164] M. Chaturvedi, "Facebook AI Reverse Engineers Deepfake Attacks," 21 June 2021. [Online]. Available: https://analyticsindiamag.com/facebook-ai-reverse-engineers-deepfake-attacks/. [Accessed 22 January 2022].

[165] J. Kahn, "Facebook says it’s made a big leap forward in detecting deepfakes," 16 June 2021. [Online]. Available: https://fortune.com/2021/06/16/facebook-detecting-deepfakes-research-michigan-state/. [Accessed 22 January 2022].

[166] Ellinika Hoaxes , "Ellinika Hoaxes," n.d. [Online]. Available: https://www.ellinikahoaxes.gr/. [Accessed 22 January 2022].

[167] M & E Studies, "Reliability Analysis in SPSS," n.d. [Online]. [Accessed 10http://www.mnestudies.com/research/reliability-analysis-spss October 2021].

[168] R. F. DeVellis and C. T. Thorpe, Scale Development: Theory and Applications, Thousand Oaks: SAFE Publications, 2021.

[169] J. M. Bland, "Statistics notes: Cronbach's alpha," 1997. [Online]. Available: https://www.bmj.com/CONTENT/314/7080/572?VARIANT=FULL-TEXT%3E. [Accessed 10 October 2021].

[170] UNL Psychology - Univerity of Nebraska, “Using SPSS Reliabilities,” n.d. [Online]. Available: https://psych.unl.edu/psycrs/statpage/alpha.pdf . [Accessed 10 October 2021].