Deepfakes and the usability of their countermeasures (Prompt 3)

Sam Barnes-Thornton (1901120)

December 2021

Introduction

The word ‘deepfake’ was first coined in November 2017, when a series of pornographic videos were posted onto Reddit with the faces of famous female actors grafted onto other actors’ bodies [1]. The name originated from the handle of the Reddit user who posted the videos, but it is now more commonly seen as the combination of ‘deep’ or machine learning and the fake nature of the media. The same technology used to produce videos like this can also be used for auditory sounds, matching someone’s voice to a user-generated script, but I will only be focusing on images and videos in this report [2]. Figure 1 is an example of the categorisation of different types of deepfakes, although it is not conclusive and is included to give an overview of the different options out there.

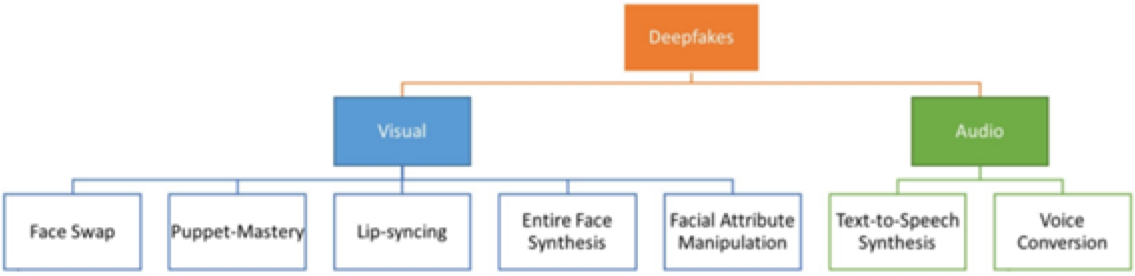


Figure 1: Categorisation of audio and visual deepfakes [3]

In this report, I will be covering the problems and security considerations that arise from this deepfake media and talking about the solutions to these. I will then discuss the usability of these solutions, and aim to produce some recommendations from this for some more usable security surrounding deepfakes.

Generation Methods

The most common technique for creating a deepfake is using a generative adversarial network (GAN), which provides a way to train a model without having access to lots of pre-existing annotated training data [4]. This is key with deepfakes as you are unlikely to have access to a large number of different images of someone’s face, especially if they are not famous. It is fairly usual for people to hear the words ‘neural network’ and stop listening as they are not interested enough to understand it, and you would think that this might deter people from even thinking about making a deepfake. However, it is very easy to find deepfake packages online with clear documentation like this python program on GitHub: `https://github.com/deepfakes/faceswap`. This means that anyone who can run a Python program, and has access to a large number of photos of someone’s face (easy with a celebrity), can create a deepfake - a scary thought.

Security concerns

Now that we know members of the public are capable of creating deepfakes, and have the tools to do it, it is important to analyse the security concerns surrounding deepfake media. The overarching danger of deepfakes is 'because of the psychological power of images, audio and video to create belief' as identified by Jacob Foster, a UCLA assistant professor of Sociology [5]. With this in mind, I will look at their different capabilities individually and the issues that arise from these. I cannot analyse every single concern that arises from deepfakes, but I have chosen several that have appeared numerous times in my research and that I believe are the most prevalent concerns.

Individual Harm

One of the things that results from good security, is an individual's privacy. Unfortunately, a big security concern arising from deepfakes is people's privacy and reputation which when revoked can, as well as damaging the person emotionally, cause long-lasting knock-on effects [6]. Sadly, 96% of deepfakes on the internet are non-consensual pornography and it doesn't need to be explained how this could affect the people who are wrongly portrayed in these videos [7]. The technology can also be used to damage people's reputation in other ways, for example it could picture a CEO saying something racist or offensive to employees. This one video could destroy their career as a company cannot cope with having it's reputation tarnished. The knock-on effects could be huge as well, possibly leading to manipulation of the stock market [8]. The key idea is that deepfakes at a low level can be used to directly harm an individual, but there can be many more consequences that follow.

In The News

One example of these other consequences can be seen in fake news which has been an important topic of interest recently, and is becoming more common everyday in a world filled with social media. In fact, it is predicted that the majority of individuals in developed economies will consume more false than true information by 2022 [9]. The problem primarily arises from social media platforms as they lack adequate regulation and their responsibilities around fake news are not fully defined [10]. Of course, deepfakes only exacerbate the problem, as they make each story that much more believable. This can lead to a common security threat from them being national security. Some terrifying examples are given in [11]:

- Politicians and other government officials could appear in locations where they were not, saying or doing horrific things that they did not.
- A fake video might depict emergency officials "announcing" an impending missile strike on Los Angeles or an emerging pandemic in New York, provoking panic and worse.
- A deep fake might falsely depict a white police officer shooting an unarmed black man while shouting racial abuse.

Obviously, these examples are extreme but they are very much possible and the impact of them has already been seen in at least a few cases like the Brexit campaign and the independence of Catalonia [10].

Biometrics

As well as the uses for impersonation in fake news, deepfakes can also be used for spoofing someone's identity in order to gain access to a system through facial recognition. Research has shown that all facial recognition systems react differently to deepfakes but some, namely Microsoft's Azure Cognition services, can be fooled up to 78% of the time by deepfakes fed into them [7]. If we look at mobile phones for an example, in 2019 facial recognition software was deployed on ninety million devices and it is expected that 90% of smartphones will have integrated biometric facial recognition hardware by 2024 [12]. This is a scary thought, as if deepfakes can be used to bypass this security, what is stopping attackers from having access to huge numbers of smartphones? It must be noted that to get a clearer picture of the efficacy of current biometric systems in detecting deepfakes, I would have to undertake far more research. However, for the purposes of this report, I will follow the evidence that most systems can be fooled by deepfakes.

Identity Theft

Most of the concerns above refer to the actions that criminals could take against targets if they are wanting to cause disruption or harm. However, deepfakes can also be used in the general workings of criminal organisations, specifically for identity theft. The current quality of deepfakes may not be quite good enough for synthetic IDs yet but Europol have found that it is possible to create so-called 'morphed' passport photos [13]. This allows a single passport to be shared by two or more individuals, and can trick both human examiners and biometric systems. They are most commonly used to fraudulently obtain documents so an element of information security is also violated here.

Something highlighted earlier was the restriction on how many photos there are available of the subject needing to be faked. A lot of the deepfakes most commonly found in the public domain depict people who are well known and therefore have a large number of photos online. However, for cases like identity theft for criminals and fooling biometrics, the subject will likely have few photos easily available on the internet. A way to combat this was developed by Russian company NTech lab, and is called FindFace, found at `https://findface.pro/`. It was originally used to find people on a social network using a photo of them, so could be utilised to find lots of photos of a particular subject that might not be easily available on the general internet. Knowing that this capability is possible, even if it is not necessarily available to everyone, shows how deepfakes can be effectively used against anyone, not just celebrities.

Existing Countermeasures

Having discussed a range of security concerns that can arise from the abuse of deepfake media, it is clear that they pose a considerable danger to society. This has been realised throughout the security industry so various different ways to combat them have been devised. In this section I will highlight the categories of solutions with some examples, and analyse the current usability of them.

The reason it is so important to analyse the usability of the solutions is because in order for a system to be secure, it must also be usable. Security and usability have often been regarded as very much competing system goals, but it is now widely accepted that this needs to change as security features have become standard components in software applications and other end-user systems [14]. Specifically in the area of deepfake solutions, usability is very important as if the detection is hard to use, then people will simply skip it and just go with the risk of media being fake.

Lots of research has been done into the best ways to make systems both secure and usable. I will be roughly following a set of guidelines provided in [14] to help structure my research. Of course, there are lots of papers providing advice in this area, but I believe this one gives a good overview and provides strong evidence for it's conclusions from a range of different sources. Regarding evidence, it is also important to note that some of the sources for these countermeasures are trying to prove their solution is the best, so I will be checking their claims against their tests.

Forensics

Throughout all of my research, this appeared to be the most common countermeasure used against deepfakes. At a high level, it involves looking at the fine characteristics of a deepfake image to see whether there are any anomalies. One example surrounds faces blinking in deepfakes. Research has shown that although the average person blinks every two to ten seconds, deepfakes have a significantly lower blinking rate [15]. This is because deepfake algorithms rely on the images available to them, and it is unlikely that a subject will have many photos available of them online with their eyes closed. Therefore, they create videos of them that do not blink as expected. One implementation of this detection method, DeepVision, was shown to be 87.5% accurate when used on videos so this demonstrates it's efficacy as a solution [16].

Along similar lines, another detection method involves exploiting missing details in the teeth area. Here, computer vision is used by first resizing facial landmarks to a generic size and then using k-means clustering to gather dark and bright clusters [17]. The bright clusters are considered to be authentic so the deepfake is rejected if less than 1% of the mouth pixels are classified as bright. This allows an easy way to detect a deepfake.

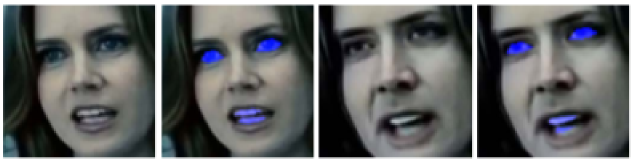


Figure 2: Face crop and clustering for detection using details in teeth [18]

The final method I will cover under the topic of forensics is multimedia forensics. This relies on analysing each phase of the image history, as different processes will leave traces (commonly known as fingerprints) in the data or metadata of an image [19]. Firstly, the acquisition of a photo can leave traces as each camera will have a certain method for processing the image and leave corresponding imperfections. Changes in these imperfections can be a clear sign of image tampering like the use of a deepfake. Compression also leaves traces, especially if it is lossy, so inconsistencies in these traces can be another example of tampering. Finally, editing an image (to make it a deepfake, for example) will modify its properties and can leave peculiar artifacts due to the processing. When combined, these points can be used to prove the unreliability of an image which is very useful in deepfake detection.

Usability

The first thing to note with the usability of deepfake forensics is the general problem that you cannot pass every single image you consume through the detection software. For example, if you receive an image to your phone you could check it in detail and pass it through software (if you had access to any). However, if you see a deepfake video on a billboard in public, it is likely to just be at a glance so implementing any kind of forensics is not possible

A second problem with usability of the solutions I have analysed, is that training is required to recognise the anomalies in images. To use the more complicated software solutions, it is likely that professional training will be required as with most digital forensic systems [20]. Unfortunately, where complicated help and guidance is required to use an application, it is likely that users will simply not use it and bypass the security [14]. As mentioned earlier, pretty much everyone sees fake media in some form or another so a lot of these people are unlikely to want to pay for professional training, making the complex forensics tools redundant in the context of general public use.

Following on from this, another guideline for usable security is that it should minimise the cognitive load whilst using a system [14]. As mentioned earlier, the forensics software is unlikely to be available to the general public so the best way for them to use forensics to detect deepfakes is by simply looking at the images to find anomalies - sometimes quite easy to spot in deepfakes like with the blinking identified earlier. However, this is not a particularly usable solution as it instantly increases cognitive load on the user because they have to analyse every image in much more detail than they otherwise would have.

A positive aspect of usability is the versatility of these solutions. Most of the methods can be used on media of any format, especially the forensics that can be performed with the naked eye. This means that users do not have to learn different methods for different sources which is a common problem with digital forensics tools [20].

Steganography

Another possible solution in the war against deepfakes is the use of digital watermarks in images. Research has been done into the best way to implement this so that recipients of images can be sure that what they receive is, or is not, a deepfake. One suggestion is to digitally sign the watermark using RSA so that only the recipient will be able to check if it is still valid or not [21]. It will be embedded into the image using steganography, and specifically inside the subject's face as that is the area we want to check for manipulation. The model in question was discovered to be 100% accurate in all testing which shows that, although simple, it is very effective. One slight problem highlighted was that including the watermark would slightly alter the pixels in the face, and therefore the facial recognition used for finding where to get the watermark. This meant that very occasionally the model would fail an authenticated image.

There are other options for implementing something similar. One has been developed by Microsoft who have a tool built into Azure that adds digital hashes and certificates to media [22]. The same hash can then be computed by the recipient of the media to check whether the image has been tampered with. There are positives and negatives to both approaches but the key difference is that the first uses a watermark specifically in the face within the image so it is checking whether that has been altered rather than the whole image. This means that it is more useful in the domain of deepfake detection than the Microsoft option, which is suited to image tampering more generally.

Usability

The most basic principle that can be inferred from all usable security research is that it should encroach on the intended use of a system as little as possible. The way this applies to deepfake detection is in the time that it takes to determine whether media is a deepfake or not, as this is likely to inhibit how quickly a user can confidently absorb the content of the media. The solution given earlier that used RSA analysed the signing and verification times of the images to try and give an insight into usability [21]. Small images had times much smaller than a second which wouldn't affect the usability of a system too much. However, the larger the number of pixels got, the longer it would take to verify the image. For example, a 1080p image with just over two million pixels would take half a second to sign but a 4K image with over eight million pixels would take roughly four seconds. As pixels numbers increase, this is a big problem as most users are unlikely to want to wait four seconds to look at an image on the internet.

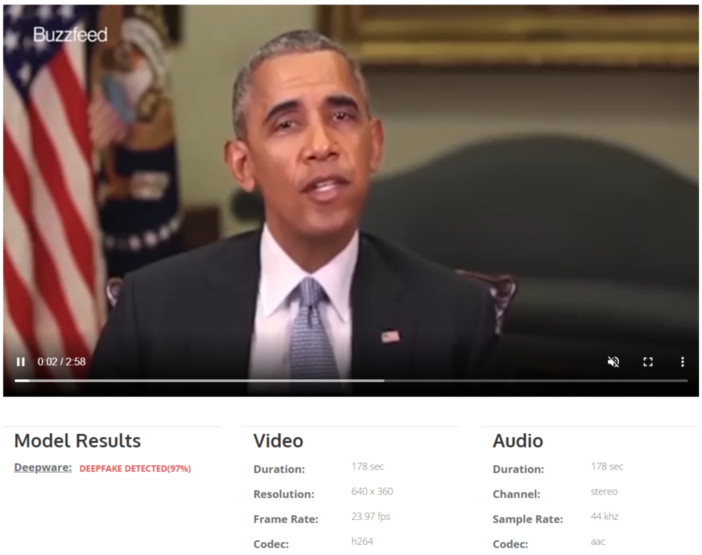
Having said that users would not want to wait, there is actually evidence to suggest that some users would prefer otherwise. A study into journalists' needs for deepfake detection found that they would prefer accuracy over a quick verdict [23]. This suggests that the needs of a specific user need to be taken account when designing deepfake detection, rather than identifying a general solution that can be implemented for all users.

Another usability issue to consider surrounding steganography in deepfakes is the guideline that the usability of cyber security should be considered early on in a design process [14]. In the case of adding signatures to images, especially in the proposed process above, it will be a feature bolted onto an existing way of consuming media which could be detrimental to the usability of the system as a whole [21]. From this, we can infer that the best way to implement any kind of detection via digital watermarks or similar would be to have a universal standard that can then be designed into new systems or new iterations of existing systems, rather than just being an add-on.

Neural Networks

A final detection method that I found in my research was the use of recurrent neural networks to perform powerful image analysis. One common implementation is to use a convolutional neural network (CNN) to extract features from frames of videos, and then use these features to train a recurrent neural network (RNN). This RNN will learn to classify if a video has been manipulated or not. There has been lots of research into how intricate differences between networks can improve detection statistics, but I have chosen one example study to give a basic overview of the efficacy of this method.

This proposed model uses a CNN as described above and then uses the features provided by this to train an RNN using a Long Short-Term Memory (LSTM) architecture [24]. The LSTM allows the network to process not only still images, but sequences of information such as videos as well. The complex inner workings of these networks are not important, but one of the finer details in how they process information is that they extract continuous sub-sequences of fixed framed length from the videos to use as input. This is because often deepfake manipulation can be present only in a small part of the video, so it is important for the networks to not only look at the video in separate chunks in order to facilitate a decision. Looking at two different examples of a model following the described structure, they both have greater than 90% accuracy in all tests which suggests that neural networks are an accurate way of detecting deepfakes [24][25]. Of course, this depends on the data that they have been tested with, but both studies clearly define the test data and explain their choices with evidence.



Model Results

Deepware: DEEPFAKE DETECTED(97%)

Video

Duration: 178 sec
Resolution: 640 x 360
Frame Rate: 23.97 fps
Codec: h264

Audio

Duration: 178 sec
Channel: stereo
Sample Rate: 44 khz
Codec: aac

Figure 3: Example of scan output on `deepware.ai`

As I mentioned earlier, there is a possibility that the evidence surrounding existing neural network solutions may not be completely trustworthy. This is because a lot of the research papers are trying to prove, or even sell, their detection method as being the best so there is a chance they might emit results that show weaknesses of their systems. However, I chose the studies used above as they are not analysing a commercial solution. They are instead researching possible improvements to existing ideas which suggests there is little or nothing in it for the authors. Consequently I believe the quality of these sources is good enough for use in this report.

Usability

I mentioned above that it was not necessary to go into the details of how the neural networks worked. The main reason for this is that it is actually too complicated to explain in the context of this paper, and that applies to the use of these methods more generally. They are seen as black-box detection methods in that the user simply inputs an image or video and gets out a decision on whether or not it is a deepfake [26]. This could cause issues surrounding deepfakes as some users may not believe a piece of technology that gives an answer with little of no evidence as to how it got there, instead opting to believe the media is real anyway. Research does show that good usability requires the ability for a user to create a mental model of a system and this is not possible with these networks [14].

However, the black-box interface style does have advantages as it does not increase the cognitive load on the user and possibly even reduces it as they don't have to check each image carefully before they can trust it. Another advantage is that methods like this can be implemented with an aesthetic and minimalist design. One example can be found at `deepware.ai`, which simply has a button to scan an image/video and then outputs a percentage chance of the media being a deepfake (seen in Figure 3). This satisfies a key guideline as some aspects of security can be complex for a novice user, so designers should aim to keep interfaces simple and reduce information overload [14].

General Usability

There are some important conclusions to be drawn from the usability of current countermeasures against deepfakes. The first is that users have different requirements from a detection system depending on how they are going to use the media that they need to check. For example, it was noted that journalists actually prefer an accurate result over speed when it comes to images they are going to publish [23]. This means that they are also more likely to want some kind of evidence rather than just a program to say yes or no, so either forensics or digital watermarking would work better for them. On the other hand, a casual user browsing social media might prefer a simple percentage score next to each image, as they are happy to believe what the computer is telling them with little or no reasoning behind it.

Another important point that can be drawn from the above solutions, is that they are all standalone systems with no integration into other software like social media. This means that users either have to go and check media separately, something novice users might not do, or the detectors have to bolted onto existing systems, going against key advice for usability in security [14]. Research does show that adding security as an afterthought is not always bad as it is often better than no security at all. It can also provide a good way to improve the front end of software if originally it was not very intuitive [27].

Recommendations

Through all my research into current countermeasures, the most notable thing I have found is that there are very few commercial solutions available. Therefore, the first suggestion I would make is for researchers to try and implement their proposed models in software that other users can utilise. The only two possibilities I have found are `deepware.ai` and Microsoft's detection tool, although the second only appears in news articles and I have not found any evidence of software that can be downloaded [22]. The first is a good example of the black-box methods that I highlighted earlier and provides mostly good usability under all the guidelines I have followed. However, an improvement I would suggest is to look at a way of integrating this into future updates to social media or news websites so that users can choose to have deepfake ratings next to images and videos, giving them an idea of whether to trust media at a glance.

One specific use-case I think it is important to highlight is biometrics. The key requirement from users here is speed as they do not want to wait a long time to unlock their phone. Therefore, the recommendation I would make here is for facial recognition systems to include a neural network detector in their process. This is because forensics were identified as being too slow for modern image sizes like two million pixels, whereas neural network classification is fast by nature. I believe this is an important improvement to biometric security systems and is not something that appears to have been talked about in existing literature.

Overall, I believe the evidence shows that neural networks are the best solution currently in terms of usability, but it is important the other solutions are not disregarded. As discussed, the needs of specific users must be taken into account so my recommendation would be to have different options on one universal ‘Deepfake Detector’. Users could choose whether they wanted an evidence-based decision using forensics or a fast decision using neural networks. This would satisfy the guideline that systems should accommodate all types of users [14].

Conclusion

This report has clearly identified a range of problems and security considerations that arise from deepfake media. It then talks about the current solutions to these and the general problems with deepfakes, critically evaluating the usability of them. I have endeavoured to make some good recommendations regarding possible future developments in this area and how to make them more usable by humans. A clear inference from the research is that there is not necessarily one good solution for all situations, so it is important to weigh up the advantages and disadvantages in each specific use-case. To add to this, the most important feature required for these solutions to be more usable in the future, is to efficiently integrate them into existing systems (social media, news pages etc.) so that users of all backgrounds can easily identify deepfake media.

References

[1] B. Paris and J. Donovan, “Deepfakes and cheap fakes,” *United States of America: Data & Society*, 2019.

[2] R. Spivak, “”deepfakes”: The newest way to commit one of the oldest crimes,” *Georgetown Law Technology Review*, vol. 3, p. 339, 2018-2019.

[3] M. Masood, M. Nawaz, K. M. Malik, A. Javed, and A. Irtaza, “Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward,” 2021.

[4] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, “Generative adversarial networks: An overview,” *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[5] S. Wolpert, “Combating deepfakes: Leading scholars to discuss doctored content and how to fight it,” Nov 2019.

[6] S. Jain and P. Jha, “Deepfakes in india: Regulation and privacy,” May 2020.

[7] S. Tariq, S. Jeon, and S. S. Woo, “Am I a real or fake celebrity? Measuring commercial face recognition web APIs under deepfake impersonation attack,” 2021.

[8] M. Westerlund, “The emergence of deepfake technology: A review,” *Technology Innovation Management Review*, vol. 9, pp. 40–53, 11/2019 2019.

[9] K. Panetta, “Gartner top strategic predictions for 2018 and beyond,” *Smarter with Gartner, October*, vol. 3, 2017.

[10] P. Fraga-Lamas and T. M. Fernández-Caramés, “Fake news, disinformation, and deepfakes: Leveraging distributed ledger technologies and blockchain to combat digital deception and counterfeit reality,” *IT Professional*, vol. 22, no. 2, pp. 53–59, 2020.

[11] R. Chesney and D. Citron, “Deep fakes: A looming crisis for national security, democracy and privacy?,” 2018.

[12] L. Pascu, “Biometric facial recognition hardware present in 90% of smartphones by 2024: Biometric update,” Jan 2020.

[13] T. Ring, “Europol: the AI hacker threat to biometrics,” *Biometric Technology Today*, vol. 2021, no. 2, pp. 9–11, 2021.

[14] J. R. C. Nurse, S. Creese, M. Goldsmith, and K. Lamberts, “Guidelines for usable cybersecurity: Past and present,” in *2011 Third International Workshop on Cyberspace Safety and Security (CSS)*, pp. 21–26, 2011.

[15] R. Chawla, “Deepfakes: How a pervert shook the world,” *International Journal of Advance Research and Development*, vol. 4, no. 6, pp. 4–8, 2019.

[16] T. Jung, S. Kim, and K. Kim, “Deepvision: Deepfakes detection using human eye blinking pattern,” *IEEE Access*, vol. 8, pp. 83144–83154, 2020.

[17] F. Matern, C. Riess, and M. Stamminger, “Exploiting visual artifacts to expose deepfakes and face manipulations,” in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 83–92, 2019.

[18] M. Albahar and J. Almalki, “Deepfakes: Threats and countermeasures systematic review,” *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 22, pp. 3242–3250, 2019.

[19] N. Gardiner, “Facial re-enactment, speech synthesis and the rise of the deepfake,” *Thesis*, 2019.

[20] H. Hibshi, T. Vidas, and L. Cranor, “Usability of forensics tools: A user study,” in *2011 Sixth International Conference on IT Security Incident Management and IT Forensics*, pp. 81–91, 2011.

[21] K. Corcoran, J. Ressler, and Y. Zhu, “Countermeasure against deepfake using steganography and facial detection,” *Journal of Computer and Communications*, vol. 9, no. 9, pp. 120–131, 2021.

[22] T. Burt, “New steps to combat disinformation,” May 2021.

[23] S. J. Sohrawardi, S. Seng, A. Chintha, B. Thai, A. Hickerson, R. Ptucha, and M. Wright, “Defaking deepfakes: Understanding journalists’ needs for deepfake detection,” in *Proceedings of the Computation+ Journalism 2020 Conference. Northeastern University*, 2020.

[24] D. Güera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, 2018.

[25] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horvath, E. Bartusiak, J. Yang, D. Guera, F. Zhu, and E. J. Delp, “Deepfakes detection with automatic face weighting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[26] L. Trinh, M. Tsang, S. Rambhatla, and Y. Liu, “Interpretable and trustworthy deepfake detection via dynamic prototypes,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1973–1983, January 2021.

[27] P. Gutmann and I. Grigg, “Security usability,” *IEEE Security Privacy*, vol. 3, no. 4, pp. 56–58, 2005.