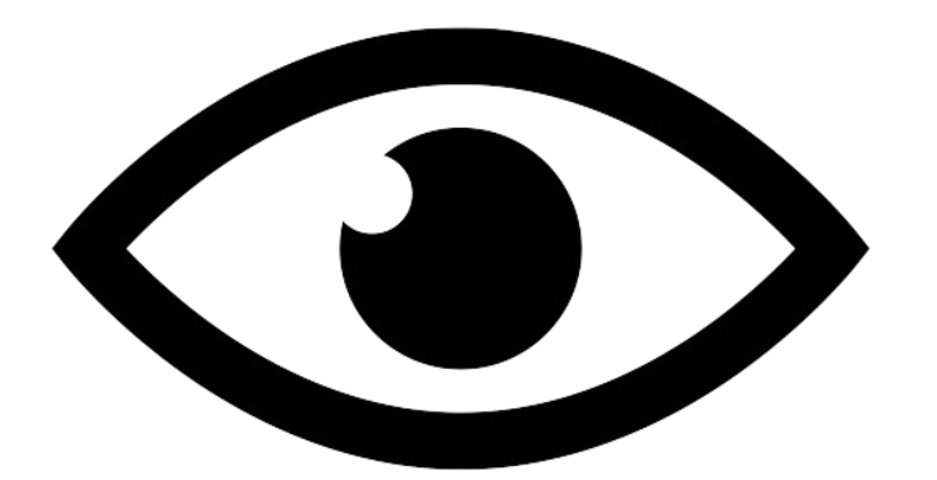
EYEdentify



Alex Buchhorn

Edward Winston

Sebastian Focas

Jonathan Willner

Luke Lewis

**Project Background**

To firstly get a grasp on why deepfake content is so rapidly advancing and becoming more easily accessible, we must understand what makes generating this content so easy. This is largely due to General Adversarial Networks (GANs), where two machine learning configurations will compete against each other in order to advance the learning of a task. It consists of a generative model that will learn to produce some type of data and a discriminative model that will learn to detect errors in the generator's output. (Goodfellow et al., 2014). The competition with each other creates an environment where both methods are improved so much so that the data generated is indistinguishable from authentic content.

Caldwell et al. (2020) conducted an analysis of crimes which were of the greatest concern in terms of AI-based criminal activity and resulted in audio and video impersonation being the highest. The use of deepfake technology can erode faith in reliable organisations and reliable forms of communication. Visual evidence can be more easily discredited if even a tiny portion of it is effectively manufactured. This can have significant repercussions for criminal investigations as well as the legitimacy of political and social institutions. A visual representation in Figure 1 of the appendix shows the comparison of all the other significant AI-generated crimes that were included in this research analysis.

It is evident that audio, video, or image-based impersonation is at the height of AI based criminal activities, the most common means in which the general populus engages with this content is via social media. Cetina & Sierra (2019) reported a large number of people using popular social media platforms had said they got some of their news from these sources: 74% of all Twitter users, 68% of Facebook users, 32% of YouTube users, and 29% of Snapchat users. These platforms have created an environment where their audiences can consume more news than ever before at any time throughout the day. It is understandable that users may be susceptible to absorbing content which may be synthetic in nature due to these platforms being riddled with gossip, personalized information, advertising and other generalised content.

**Gap in the Market**

Through our independent research we have identified that there is a gap in the market in detecting deepfakes for social media and general internet users. Social media usage in particular is done on phone applications as well as on web browsers and most sites don’t have deepfake detection built into them. Because of this its left to the end users to decipher whether or not images are deep faked or not. Research has been done on detecting deepfakes and a series of detection tools were developed. However, these tools are not readily available to the end users. This is clear when you try to obtain a method of detecting deepfake images and video on an apple phone. There aren’t any apps with the capability to detect deepfake images and videos. The same can be said about android and chrome/ firefox web extensions. While looking to see what was on the market already there was a major lack of detection tools and those that did exist lacked significant features that would make them worthy of using. Figure 2 is an example of how current products in the market are able to meet the standards required to make an impact in deep fake detection software analysis.

**Google Chrome and Firefox Extensions**

Duck Duck Goose

* Compatible with images only
* Relatively long scan time

Deepfake, GAN

* Only works for GAN generated material
* No video or audio compatibility

AI Extension for Youtube

* Only works for videos but not audio

Firefox

* No scan function
* Only compatible with GAN generated material

We were unable to find research on why there is a lack of apps for detecting deep fakes, but it is likely due to the long time needed to be approved for an app which results in detection being worked around. The other reason is likely due to the computing power needed for detection which older phones do not have and only some of the newer phones have. On google chrome and firefox we were able to find a few deepfake detection applications. 2 out of the 4 didn’t work at all. The first one was called duck duck goose. It could only scan images and short videos but took on average 90-120 seconds per image and 600 seconds for a 10 second video. Using thispersondoesnotexist.com to generate faces it was accurate 88% of the time wish a sample size of 50. Fake profile detector was very inaccurate in testing. It was limited to only being usable on GAN but had a high rate of false positives; even when the same image was used twice it would have different results. The third detection tool could only be applied to YouTube videos. This was not its only limitation as it could not detect deep faked images within a video. This was found by using a deep faked image and uploading it as a 30second video which was not detected as a deep faked video or image. The last tool was unable to scan images even though it was listed as being able to. This was likely due to it being listed as a prototype on the firefox extensions site. All four of these products were not available for a phone application.

Due to the wide range of crimes perpetrated through the use of deepfakes, several notable theories support the necessity of our proposed prototype. Our prototype is built to detect deepfakes and as a result can be used to prevent crimes on a number of targets ranging from large scale organisations, corporations, and governments to small scale individuals and stakeholders (Drew, 2020). As a result, the need to verify content remains a critical issue. These crimes can have a high cost to businesses such as one case that lost a business 35 million, as well as high cost to life, such as the attempted coup in Gabon in 2019 (Cahlan, 2020). Rather than focusing on one specific area of crime, our solution has a wide range of applicability which are justified by the Routine Activity criminological theory.

Routine Activity Theory is commonly applied to cybercrime to explain victimization and criminality (Kigerl, 2021). This approach posits that victimization is caused by a combination of three key areas converging in time or space. These areas are a motivated offender, which is an inherent factor, a suitable target, which can be analysed through the acronym VIVA (Value, Inertia, Visibility, and Access), and capable guardians, such as our proposed deepfake detection tool (Kigerl, 2021). As a result, eydentifier needs to address the four areas of a suitable target. The value of a target cannot be minimised, as attackers often resort to either targeting key figures with a high value such as CEOs, or a wide range of victims with a cumulative high value (Drew, 2020). Inertia relates to the level of barriers and risks present to the offender. As a gap in the market remains to address deep fakes, there is currently little to no barriers present for this type of crime, regardless of the risk present. By providing not only the ability to detect deepfakes, but report them, both the risk and barriers the offender faces can be raised. Visibility refers to the ease of access the offender has to commit the crime. As there are no current mitigation strategies to reduce the levels of victimisation for these users, the visibility for offenders remains high allowing offenders to reach a wide audience through a number of methods such as social media and targeted attacks (Drew, 2020). By implementing deep fake detection, the number of unprotected users can be reduced, lowering the visibility of the target. Access refers to the ease of offending. As a result of current trends in online spaces, an estimated 81% of the Australian population uses social media (Cristina, 2023). Access to social media remains unrestricted, allowing bots and automated services to proliferate deepfakes amongst these apps and reach a wide audience. While access will not be mitigated through our prototype, by allowing users to verify the images they see, the number of victims can be reduced. By introducing a capable guardian like Eyedentifier, the prototype has the potential to minimise the widespread proliferation of deepfakes, reducing the accessibility and visibility of victims, and raising the barriers and risks present to offenders.

**Prototype**

The solution consists of two prototypes: a web browser extension and a mobile application. In February 2022, internet traffic in Australia showed that 55.41% came from desktop users and 39.88% from mobile devices (Statista, 2022). Excluding one of these groups, would have significantly hindered the reach of our solution, thus making it vital that two separate prototypes were devised, one for each platform. Although they each cater to different target audiences; desktop users and mobile users, both are designed specifically for public use. Notable stakeholders include users of dating apps and social media platforms, who are wanting to alleviate anonymity and add a level of assurance and verification, mitigating deepfake scams. Beyond this, celebrities, politicians and other high-profile figures who are most prone to identity theft through deepfakes would also heavily benefit. One way to evaluate the solution, would be to measure the number of downloads, offering insight into the reach of the product and its word of mouth. To better evaluate the functionality, further scans of deepfake materials with a count of false positive test results, would prove vital. Lastly, a measure of reports made within our app, will outline the extent at which the prototype is being used. Where the project presents its limitations, is its inability to scan video and audio deepfakes. Given the small-scale development team and resources available, incorporating such an exhaustive feature was not feasible within the given time frame. Additionally, given the nature of how mobile applications function, automatic scans of media sources are not yet possible. Users are required to upload images manually to the app, whereas the browser extension performs background scans. Despite these limitations, both protypes succeed in several areas, and present an immediate need for appropriate measures to counter the increasing risk of deepfake crime.

The largest risk present to this prototype is the security of the user's data on where their data is stored. Each country has its own requirements on how data can be stored safely and how much personal identifiable information can be stored (such as name, date of birth, and region) (Komaitis, 2017). To account for data privacy and data security concerns, the data will be localised and hosted on servers within Australia. As this prototype is intended for an international audience, the security and privacy setting must comply with the European General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) (Komaitis, 2017). As these are the strictest regulations on what data can be stored by businesses and what users can access of their data, by adhering to these rules and regulations, the data privacy and security concerns present within this prototype can be mitigated.

A key aspect of this prototype is the reporting feature. This has been included for users to report detected deepfakes directly to the team. This allows us to run further verification checks to ensure that a false positive has not been created. If the image is verified as a deepfake, then further actions can be taken by Eyedentifier to contact the host to remove the image, or in extreme cases report to the police. Meta (2022), the parent company for Facebook and Instagram has created a new policy where they will only remove images under two conditions: It is edited in ways that the average person would not detect and would mislead the user, or if the image is created by AI to appear authentic.

**Team Reflection**

The initial project development team consisted of Alex Buchhorn, Edward Winston, Sebastian Focas, Jonathan Willner, and Luke Lewis. The entire team worked on developing the block framework and the initial project pitch. Alex Buchhorn Created the key UI elements such as the logo and confidence indicators. A detailed mock-up of the browser extension was completed by Jonathan Willner and Edward Winston, and a mock-up the mobile app by Luke Lewis and Sebastian Focas, with both mock-ups having a final delivery. Finally, testing, final edits and displaying wireframes was completed and performed by all team members on the 4th of October. This will mark project completion on the 11th of October. In addition to delegated tasks, each team member acted as consultants to review and assist other members work where necessary. Each deliverable was completed on time with no significant delays occurring during production of the project. Details of all project tasks and deliverables are provided in the timeline shown in figure 17 of the appendix. Team communications were facilitated through 3-hour, weekly, in-person meetings where key deliverables and deadlines were discussed, and tasks were delegated. Additional communications were conducted through a Microsoft Teams channel created at the beginning of project development, as well as through individual communications between team members.

Project risks regarding team contribution include team members being unable to complete tasks prior to the deadline, such as project wireframes, testing, and market research. Another risk exists in essential functions of the prototype being non-functional, such as the ability to accurately detect deep-faked content. During development of the project none of these risks came to fruition and all essential functions were complete.

**References**

Cahlan, S. (2020). Analysis | how misinformation helped spark an attempted coup in Gabon. The Washington Post. https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/

Caldwell, M., Andrews, J. T., Tanay, T., & Griffin, L. D. (2020). AI-enabled future crime. Crime Science, 9(1), 1-13. https://doi.org/10.1186/s40163-020-00123-8

Cristina. (20239). *Digital 2023 Australia: 1 in 3 Australians use social networks for Brand Research*. We Are Social Australia. https://wearesocial.com/au/blog/2023/02/digital-2023-australia-1-in-3-australians-use-social-networks-for-brand-research/

Drew, J. M. (2020). A study of cybercrime victimisation and prevention: exploring the use of online crime prevention behaviours and strategies. Journal of Criminological Research, Policy and Practice, 6(1), 17–33. https://doi.org/10.1108/JCRPP-12-2019-0070

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems.

Leukfeldt, E. R., & Yar, M. (2016). Applying routine activity theory to cybercrime: A theoretical and empirical analysis. Deviant Behavior, 37(3), 263-280.

Cetina Presuel, R., & Sierra, J. M. M. (2019). Algorithms and the news: social media platforms as news publishers and distributors. Cetina Presuel, R., & Martínez Sierra, J.(2019). Algorithms and the News: Social Media Platforms as News Publishers and Distributors. Revista De Comunicación, 18(2), 261-285.

Kigerl, A. (2021). Routine activity theory and malware, fraud, and spam at the national level. Crime, Law and Social Change : An Interdisciplinary Journal, 76(2), 109–130. <https://doi.org/10.1007/s10611-021-09957-y>

Komaitis, K. (2017). The ‘wicked problem’ of data localisation. *Journal of Cyber Policy*, *2*(3), 355–365. https://doi.org/10.1080/23738871.2017.1402942

Meta. (2020) *Enforcing against manipulated media*. . <https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/>

**Appendix**

*Figure 1: Crimes in the upper right corner are considered extremely harmful, most profitable, and hard to combat against. Crimes in the lower left are considered the least profitable and easiest to defeat (Caldwell, et al., 2020).*

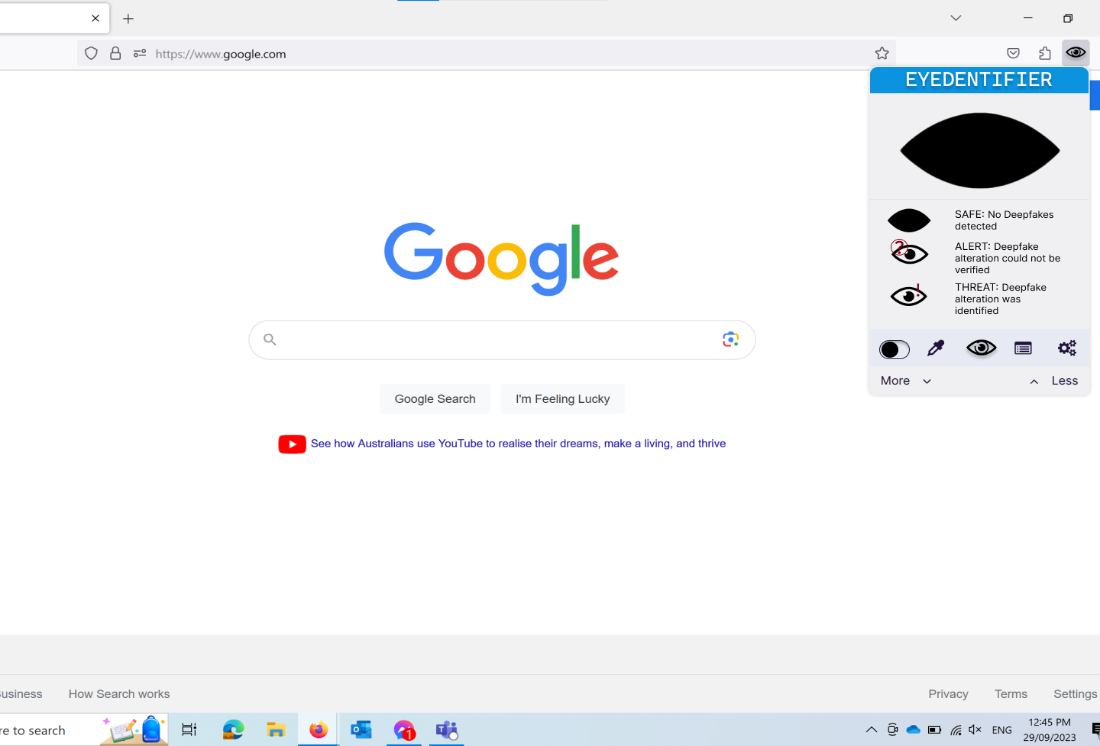
A diagram of a diagram

Description automatically generated with medium confidence

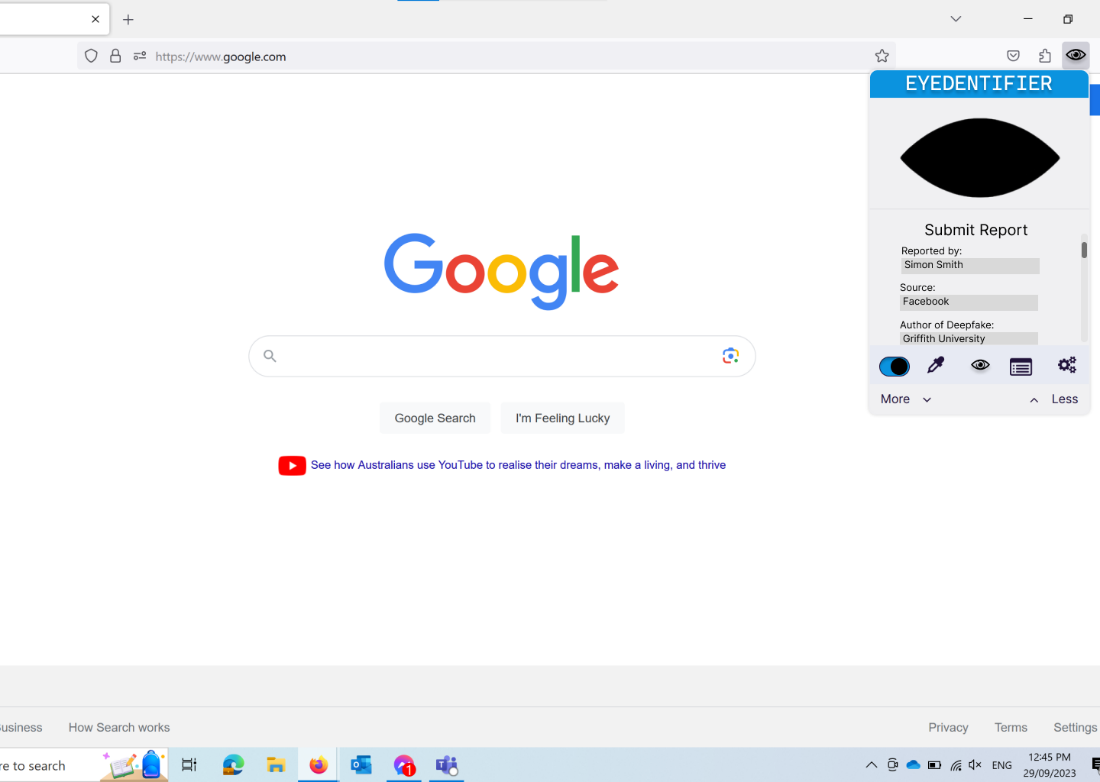
*Figure 2: Existing Deepfake detection software*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Product** | **Quick Scan** | **Mobile App** | **Scan for all Types of Deepfake Images** | **Self-Upload Images** | **Link** |
| Deepfake Proof (Very Long Scan Time) | No | No | Yes | No | <https://chrome.google.com/webstore/detail/deepfakeproof/ehjldchkbnfkmicpofahcghimhkkpngo?hl=en> |
| Fake Profile Detector (GAN only) | Yes  (But Inaccurate) | No | No | No | <https://chrome.google.com/webstore/detail/fake-profile-detector-dee/jbpcgcnnhmjmajjkgdaogpgefbnokpcc?hl=en> |
| Deep Fake Detection (Videos) | Yes  (Scan time relatively efficient for videos) | No | No | No | <https://chrome.google.com/webstore/detail/deepfake-detection/gbokgdgbgfcdlbnonmlajiapbcpkdgnk?hl=en> |
| Deepfake Detector (Firefox) | No  (Didn’t Scan) | No | No | No | <https://addons.mozilla.org/en-US/firefox/addon/deepfake-detector/?utm_source=addons.mozilla.org&utm_medium=referral&utm_content=search> |

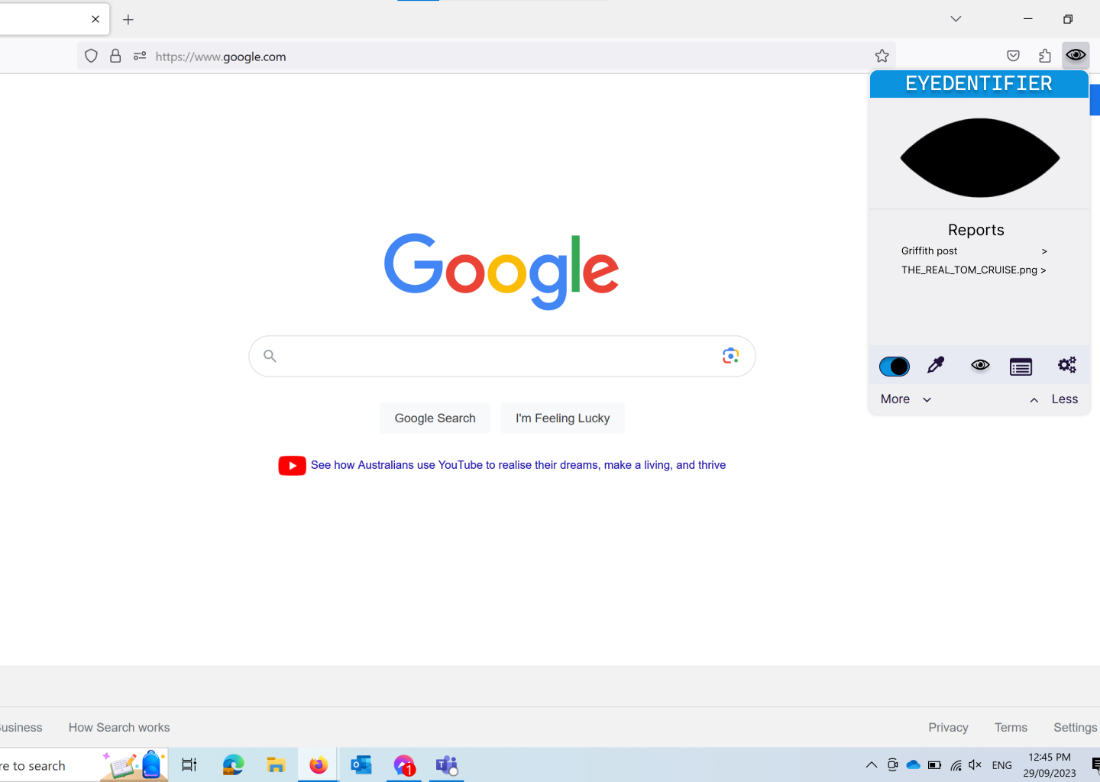
*Figure 3: Logo Schema*



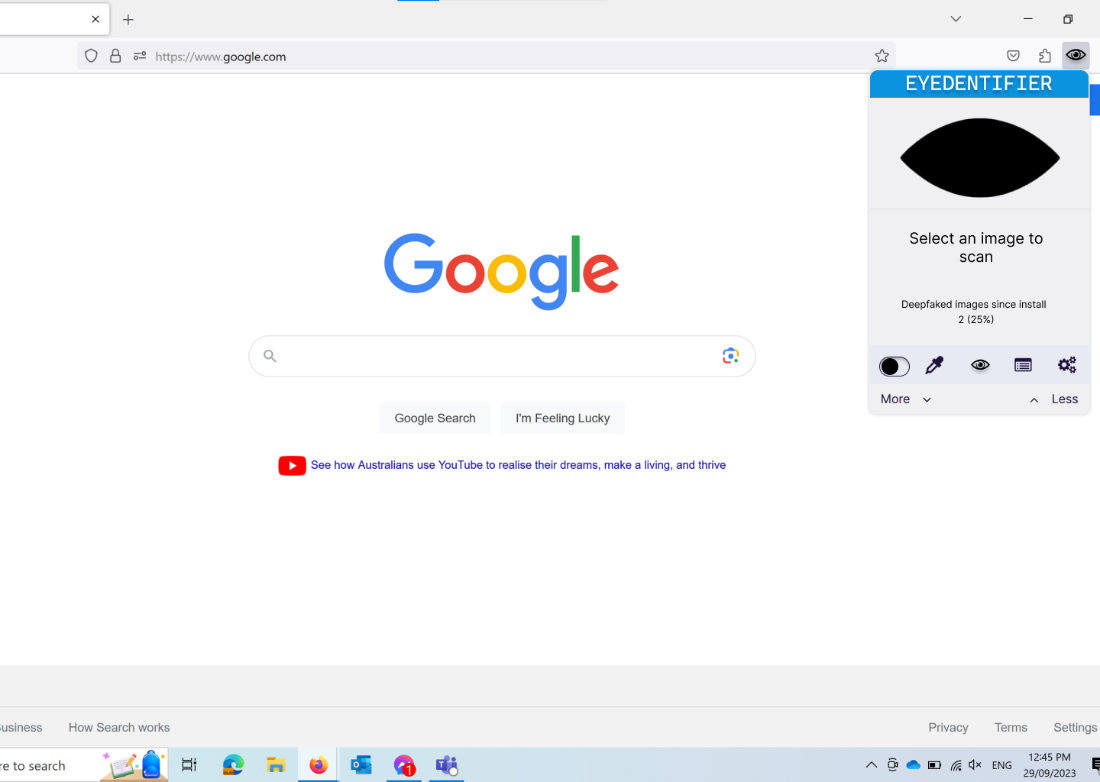
*Figure 4: Reporting page*



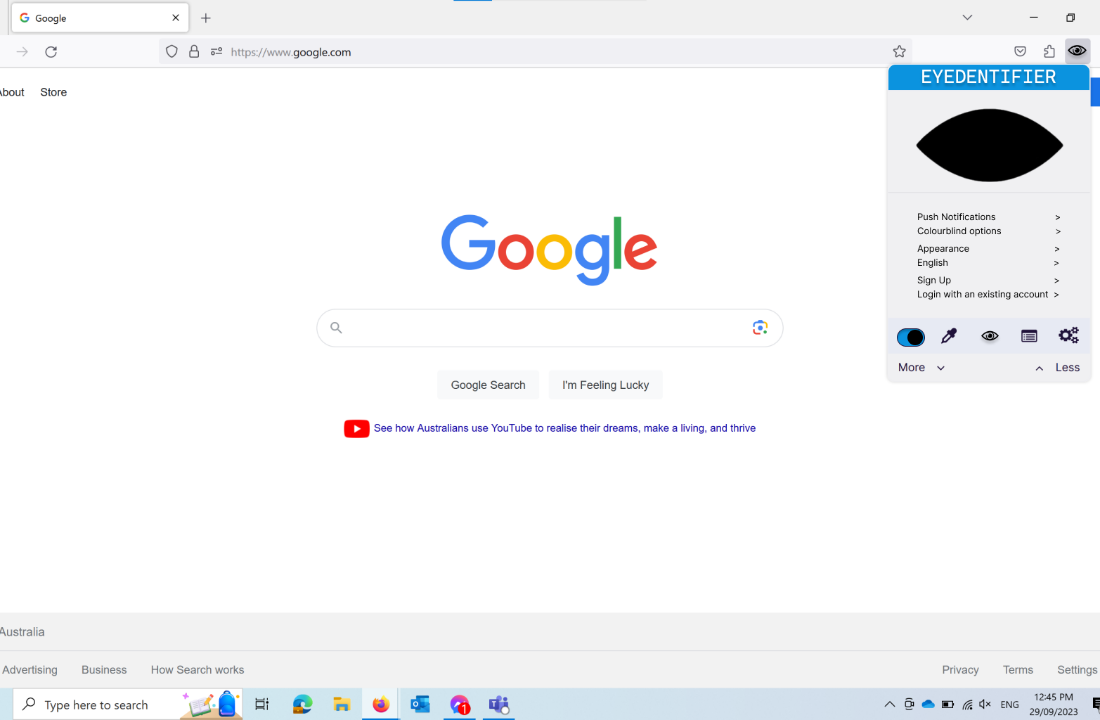
*Figure 5: Previous reports page*



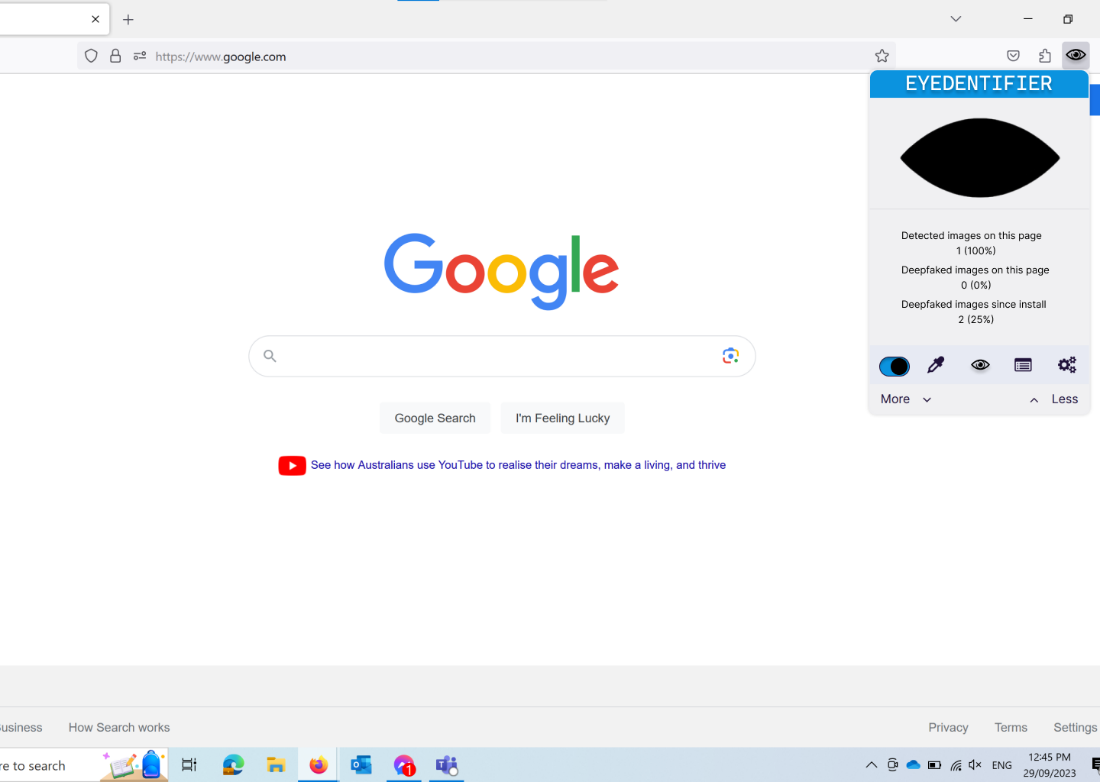
*Figure 6: Manually select an image*



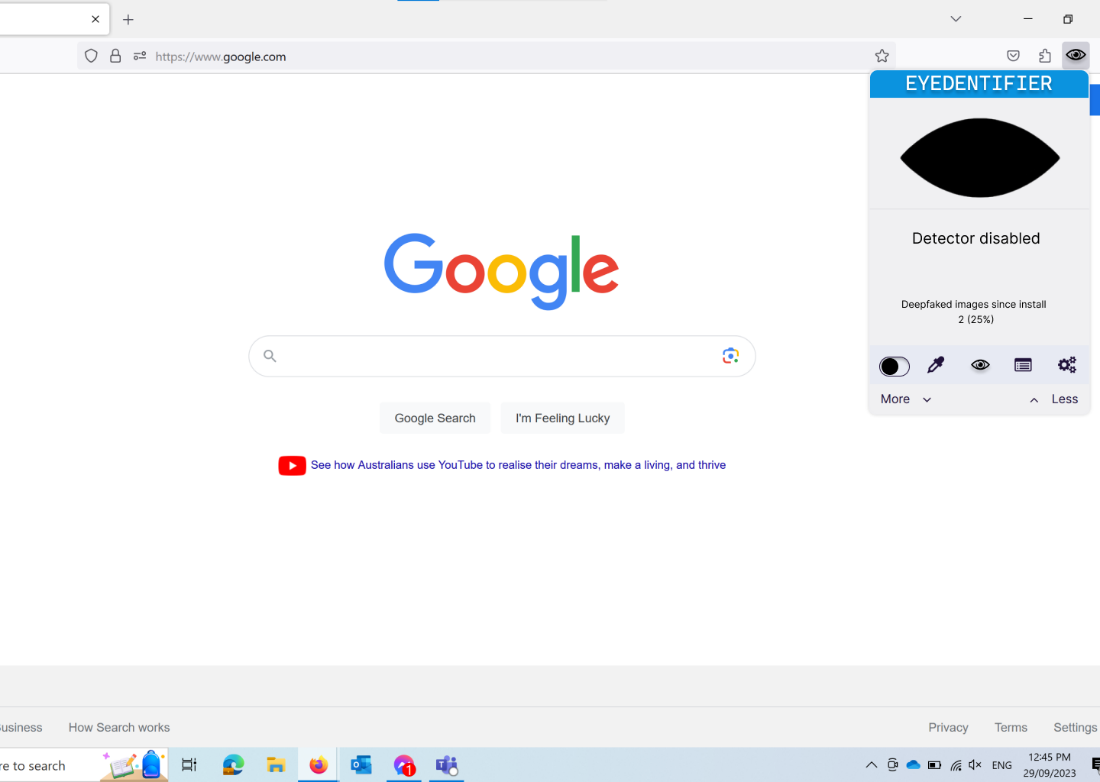
*Figure 7: Settings page*

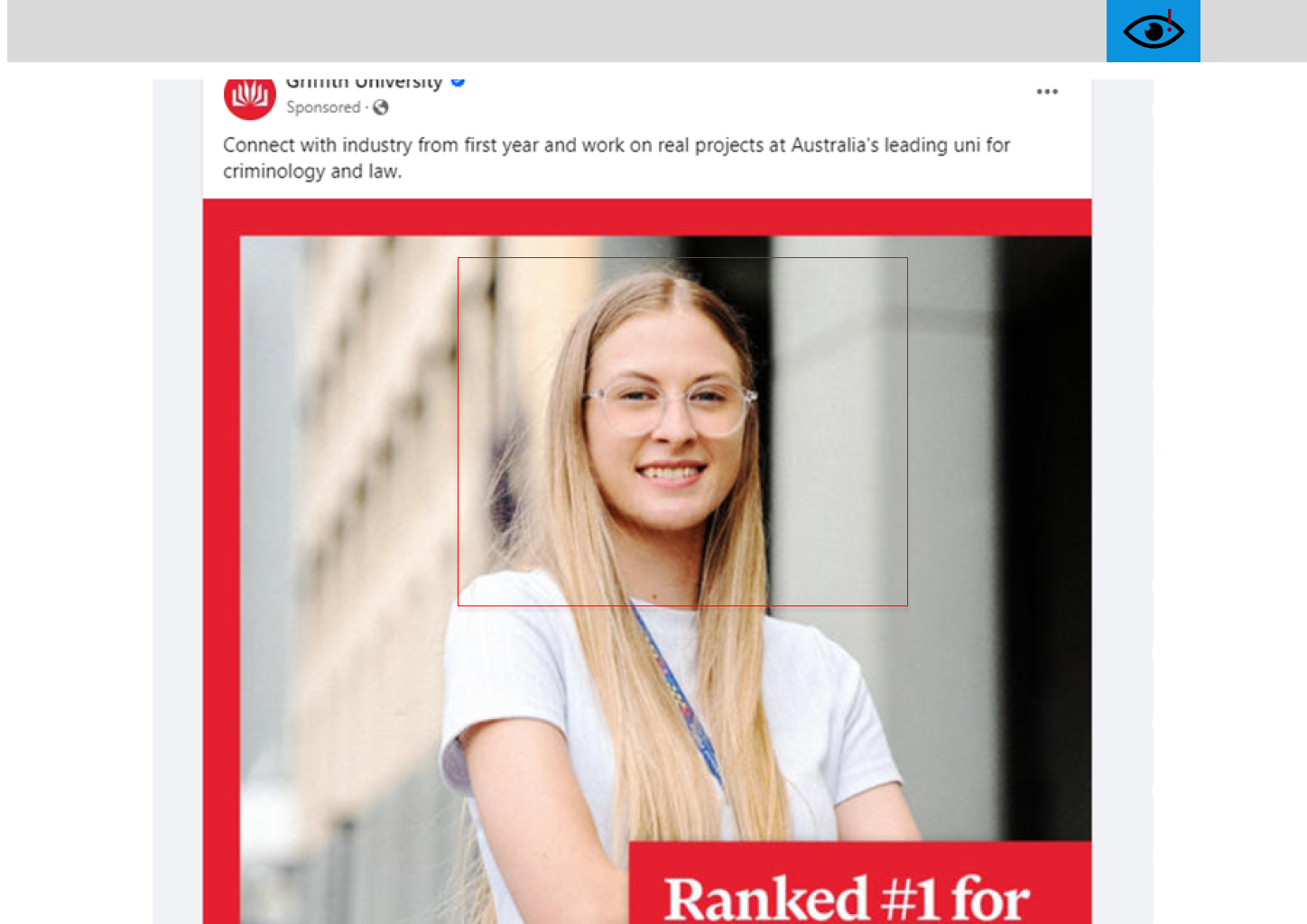


*Figure 8: Results page*

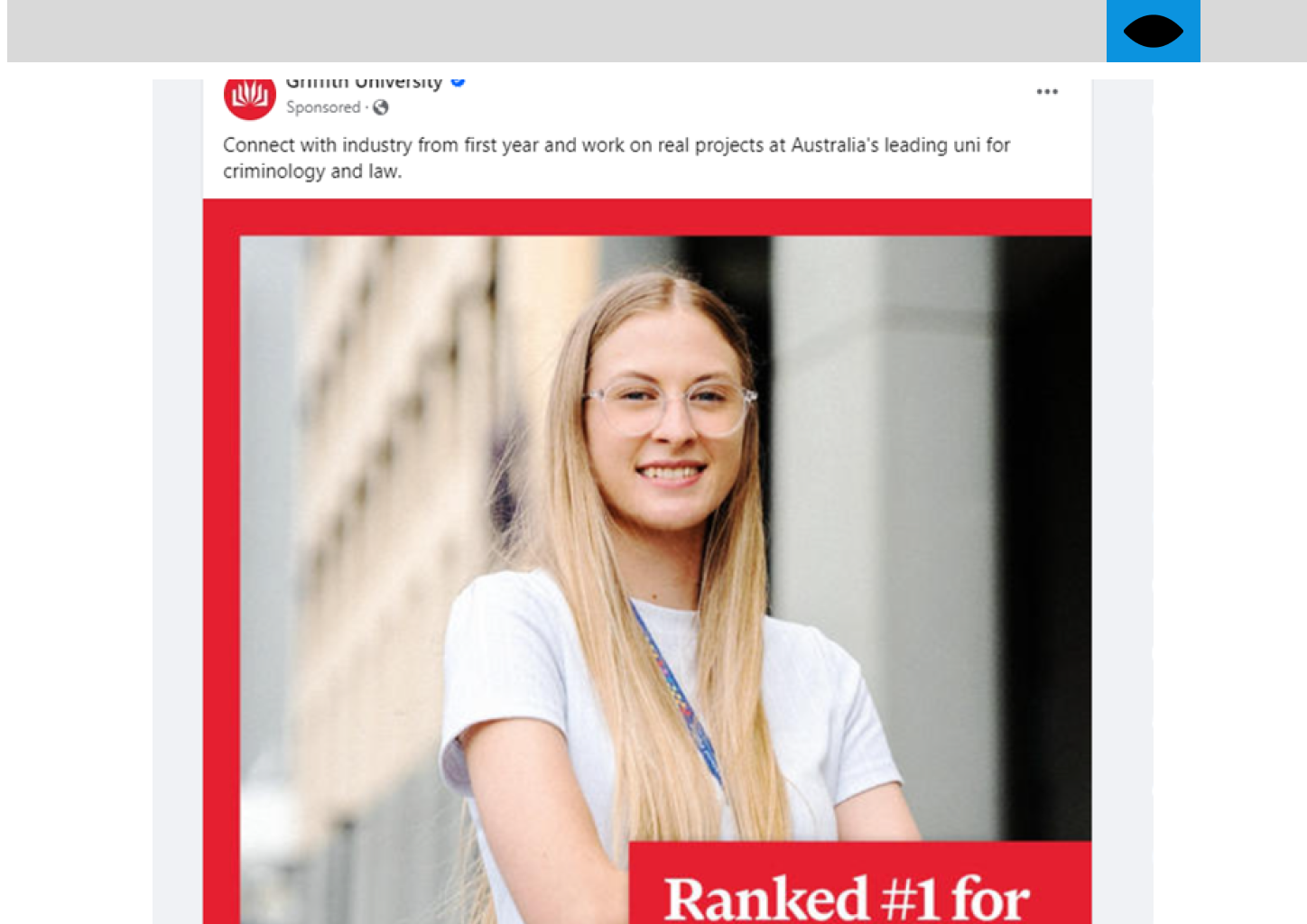


*Figure 9: Disabled landing page*

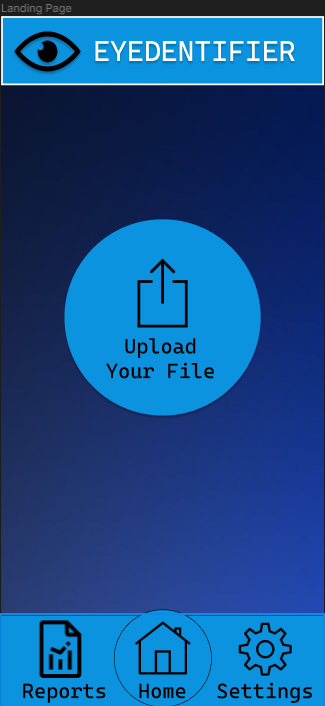
  
*Figure 10: Scanner detected deepfake*



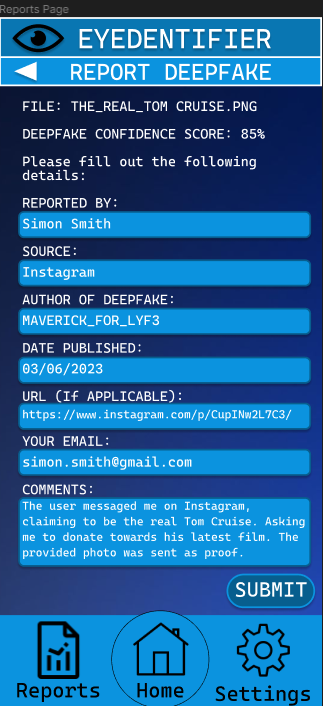
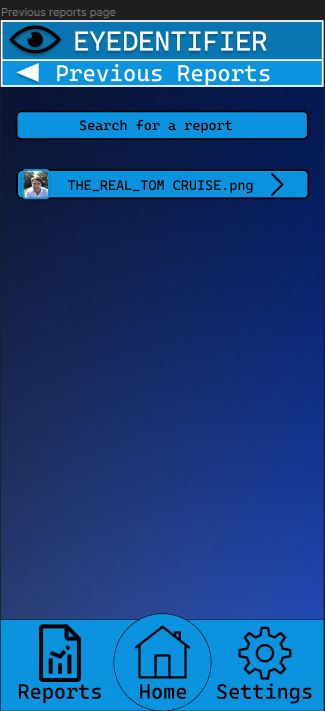
*Figure 11: Active scanner did not find deepfake*



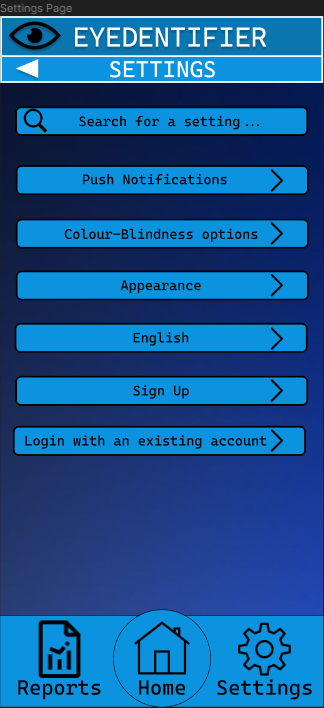
*Figures 12 & 13: Landing page and result page*



*Figures 14 & 15: Reports & previous reports page*



*Figure 16: Settings page*



*Figure 17: Project deliverables and timeline*

|  |  |  |  |
| --- | --- | --- | --- |
| Project Task | Proposed Deadline | Completed Date | Reflection/  comments |
| Project start | N/A | N/A | Project commencement 12 weeks prior to completion |
| Block framework | 30/8/23 | 30/8/23 | Finalised during weekly meeting. |
| UI logo elements created | 6/9/23 | 6/9/23 | Delivered during weekly meeting. |
| Browser extension mockup | 27/9/23 | 27/9/23 | Delivered prior to meeting to prepare project pitch. |
| App extension mockup | 27/9/23 | 27/9/23 | Delivered prior to meeting to prepare project pitch. |
| Testing | 4/10/23 | 4/10/23 | Conducted on delivery date. |
| Final edits and showcasing wireframes | 4/10/23 | 4/10/23 | Performed directly after user testing. |
| Project finish | 11/10/23 | 10/10/23 | Project Complete. |