THE BROKEN SHIELD - FACE RECOGNITION, ADVERSARIAL ATTACKS, AND THE DANGERS OF OFFERING IMPERFECT TOOLS

K. Gretchen Greene *
Assembly Fellow, Berkman Klein Center for Internet & Society at Harvard University
Fellow, Belfer Center Harvard Kennedy School
Senior Advisor, The Hastings Center
Partner, Baker Thomas Oakley Greene PLLC
gretchen@bakerthomaslaw.com

Thomas Miano
Assembly Fellow, Berkman Klein Center for Internet & Society at Harvard University
Research Data Scientist, RTI International
tmiano@rti.org

Daniel Pedraza
Assembly Fellow, Berkman Klein Center for Internet & Society at Harvard University
Trust & Safety, Berbix Inc.
daniel@berbix.com

Kira Hessekiel
J.D. Candidate, Harvard Law School
khessekiel@cyber.harvard.edu

ABSTRACT

In this article we survey adversarial attack tools for face detection and recognition models (FDRMs[1]) and present and analyze a case study of risks associated with FDRMs and adversarial tools. We develop a framework for user-centered design of adversarial tools and draw lessons on risk disclosure from U.S. product liability, consumer protection, and negligence law. Combining these, we develop a set of concrete recommendations to help FDRM adversarial tool makers decrease user harm through better communication around risk.

1 INTRODUCTION

Adversarial attacks have been presented as a tool for avoiding FDRM surveillance since 2018, with pioneering work by equalAIs (Pedraza et al., 2018; Miano, 2018; Adjodah et al., 2018), an MIT/Harvard cross-disciplinary team. Use and implementation of FDRMs, previously available only to technology corporations and research scientists, are now ubiquitous with applications like unlocking phones, exam proctoring, customer service, and detective work. They are also increasingly available to the general public. Founded in 2017, Clearview AI created a tool that scrapes the internet for information about individuals using images of their faces and made it available to law enforcement and others (Hill, 2020). PimEyes, commercialized in 2019 (Wakefield, 2020), offers a similar service that was used by independent citizens to search for clues after the January 6 U.S. Capitol riot (Hines, 2021). Several big technology companies offer FDRM APIs as part of their cloud services Microsoft Azure (2021); Google Cloud (2021); Kairos (2021). Building FDRMs has

*https://cyber.harvard.edu/people/gretchen-greene

[1]In our article we use the following definitions: face detection is the process of locating a face in an image; and face recognition is the process of predicting the identity of a face in an image.

also become much easier in recent years. On-demand compute services have become common, face detection algorithms, code, and training sets are freely available, and technical knowledge is much more widespread. As FDRMs grow in popularity and usage, we should be thinking about what tools and levers of control individuals (could) have to avoid being seen, and we must consider whether they have the potential to cause harm to the very individuals they were meant to protect.

In this paper, we briefly survey the history of FDRM attack tools and prototypes, discuss the claims made about their efficacy and robustness, and consider the harms these tools could cause individuals, primarily the harm of failure when a user expects the tool to work.

We draw lessons from U.S. product liability, consumer protection, and negligence law. We use these areas of the law as points of reference for considering society's expectations about: (1) the need for warnings and truth; (2) the duty to create user understanding; and (3) reasonableness and compliance with community and industry norms and standards. We apply these lessons and analyze a use case scenario to create a framework and a concrete set of recommendations for avoiding harm when creating and talking about FDRM adversarial tools.

2 FDRM ADVERSARIAL TOOLS AVAILABLE TO THE PUBLIC

With the rise in popularity of adversarial model research, many adversarial tools that change a photograph have become available in the realm of FDRMs. In early 2018, a group of fellows at MIT and Harvard released a tool called equalAIs, which prevented face detection against several major FDRM APIs and was available to the public as a web tool. The team also proposed adding steganographic messages and visible watermarks to allow users to communicate that they do not consent to their photos being processed by FDRMs, which were implemented in early 2021. Later in 2018, researchers at the University of Chicago released an adversarial tool, Fawkes, which is described as being intended for personal privacy protection and academic research (Shan et al., 2020) and is described as being "100% effective against state-of-the-art facial recognition models" (Microsoft Azure Face AI, Amazon Rekognition, and Face++), with a caveat identified in 2021 that their efficacy against Azure has lowered (Shan et al., 2021). More recently, a group of researchers have released LowKey[2], an FDRM adversarial web tool that provides users with a toggle bar to adjust the level of perturbation in the image, and changes the level of reported reliability of the tool. LowKey makes advancements in the space by introducing a novel architecture to build the tool with expanded testing and evaluation (Cherepanova et al., 2021).

3 THE HITMAN'S QUARRY AND THE HARM OF MISUNDERSTANDING THE RISK OF FAILURE

There are many reasons why someone might want to avoid being identified in a photograph or might not want to live in a society where it is very easy for strangers to access vast amounts of information about them. We consider a single case study for the use of FDRMs and adversarial attacks, and use the case study to establish a framework for analyzing harms.

3.1 THE HITMAN'S QUARRY

Scenario/fact pattern[3]. A federally protected witness (the user) with a new identity gets tracked down and killed by a member of the gang (the pursuer) they testified against who:

1. Uses an FDRM product, "Clairvoyance," to search the internet for the user's face
2. Finds the user's picture, which was recently posted by the user on a social media site
3. Notes clues on the site about the user's new identity and physical location
4. Finds the user and kills them

[2]Tool available at: https:// lowkey.umiacs.umd.edu.

[3]There are other very similar fact patterns we might have used. Victims of domestic violence fleeing their abusers frequently seek to change their names, addresses, and identities. (Driskell, 2009)

Harm/loss to the person whose face is in the image. Death

Using an adversarial attack tool could avoid or reduce the loss. Clairvoyance would have failed to find the new picture on social media if the user had processed it with a successful adversarial attack filter before posting it.

User alternatives to an adversarial attack tool. The user could have avoided harm by not posting the picture at all. Depending on the social media platform's available security settings, and the pursuer's determination, the user may have been able to avoid harm by enabling settings that would prevent Clairvoyance from accessing the photograph or other user metadata.

Both the adversarial attack's failure and the user's poor understanding of the risk of failure are dangerous. If the user falsely believes adversarial attacks are sufficient protection, then they may rely on them and fail to take other actions that could complement or be better than adversarial attacks as protection, incurring the loss that successful face recognition causes if the adversarial attack fails.

3.2 Framework for User-Centered Design of Adversarial Tools

We can expand our analysis in the Hitman's Quarry to create a framework for working through any scenario/fact pattern to identify risks and risk mitigation strategies related to FDRM adversarial attacks from a user centered-design perspective.

1. **FDRM Harm.** How can FDRM harm the person in the image?
2. **AT Defense.** How could that person avoid the harm with an adversarial tool?
3. **Alternate Defense.** How else could that person avoid the harm?
4. **AT Harm.** How could that person be harmed when the adversarial tool fails?
5. **AT Harm Avoidance.** How can that person avoid harm from adversarial tool failure?
6. **Additional Risks.** What additional risks of harm from an adversarial tool failure are there for vulnerable groups in this scenario?

Empower users to evaluate the risk of tool failure. While tool makers can help users avoid harm by improving the tool itself, another (perhaps much more important) way tool makers can reduce harm, is to ensure that users are empowered to adequately evaluate the likelihood and consequences of tool failure. For example, even if an adversarial tool is known to work at the time a photograph is posted, there is still a substantial risk of future failure as face recognition technology improves over time, a risk the user is unlikely to be aware of[4]. The difference between an attack and a defense is also important to note, if tools are being used as a defense against FDRM harms. In many applications, an attack only needs to work once, while a defense needs to work every time (Garg, 2020).

4 Lessons for FDRM adversarial tools from product liability, consumer protection, and negligence law

The idea that tool makers should give users appropriate information to help users evaluate risk and avoid harm is hardly new. We see it in the law of product liability, with fainter echoes in consumer protection law and perhaps even common law negligence. We look at these three areas of the law to draw lessons about (1) the need for warnings and truth; (2) the duty to create user understanding; and (3) reasonableness and compliance with community and industry norms and standards.

[4]The major face recognition APIs are not static. Engineers are continuously improving the models and training sets. An adversarial attack that isn't updated should be expected to lose effectiveness over time, even if the face recognition engineers aren't actively trying to counter the attack (and it will happen much faster if they are). A user's face on an adversarial attack processed photograph may be undetectable by every existing face detection program in the world today when the user puts the photograph somewhere on the internet where they cannot control its distribution or pull it back. And someday in the future, maybe tomorrow, the face will suddenly become visible to automated detection. CV Dazzle (Harvey, 2021), a project that uses makeup and hairstyle designs to avoid detection by FDRMs, has a prominent warning on its website about exactly this kind of risk.

4.1 PRODUCT LIABILITY FAILURE TO WARN

Tort law recognizes a category of liability where a seller's product presents a serious risk to the consumer that they could not reasonably anticipate, otherwise known as failure to warn. If the consumer uses the product in an expected way and is harmed in a manner that an average person would not foresee, the seller is liable even if they did what would normally be considered enough to protect against negligence liability, and even if the consumer did not purchase or enter into a contract with the seller[5]. In the world of cybersecurity, these claims have been thought to be preempted by federal cybersecurity laws like CISA[6]. For designers of adversarial tools, failure to warn about the limitations and risks of the tool might give rise to this sort of liability, since providers offer and users seek the product for protection against the harms that might come as a result of FDRMs, while users may not reasonably anticipate the limitations of protection that the tool can provide.

4.2 CONSUMER PROTECTION

The U.S. federal government and all fifty states have legal regimes designed to protect consumers from unscrupulous commercial practices and that give government entities the right to pursue violating companies for civil and criminal enforcement on behalf of consumers. Many state statutes are modeled after the foundational Federal Trade Commission Act, which created the FTC and gave it the power to enforce a prohibition on "unfair or deceptive acts or practices in or affecting commerce."[7] The FTC and the states interpret these powers expansively and seek to hold entities accountable for the promises they make to consumers about the benefits and capabilities of their offerings. The FTC in particular has been active in the technology sector in investigating assurances made by companies regarding their cybersecurity practices, and even their compliance with other nations' data privacy laws (like the GDPR or the E.U.-U.S. Privacy Shield Framework) and bringing enforcement actions if their descriptions do not match reality. In general, these laws seek to ensure that all businesses are forthright about benefits and limitations of their products.

4.3 NEGLIGENCE

Negligence is the common law tradition's means of rectifying harm when the actor does not act with an intent to harm. It is not necessarily tied to a particular statute or situation - in its core formulation, negligence means "conduct which falls below the standard established by law for the protection of others against unreasonable risk of harm."[8] The conduct may be an affirmative step taken by the actor, but it can also be a failure to do something it could reasonably do. In general, an unreasonable risk is one where the cost of the harm caused, when factored by the probability that the harm would occur, is greater than the cost of the preventative steps the actor could take. There are factors to consider beyond the pure cost balancing on the facts of the case. If the government enacts a law or regulation that categorizes some activity as "unreasonable," that is usually sufficient to make conduct unreasonable (absent some larger constitutional or administrative defect with the law itself). It is also important to be aware of more general practices around what is considered statutorily "unreasonable" in commerce and tech. Negligence law often relies on evidence of industry standards and community customs, whether formally written down through an industry standard-setting organization, or informally surmised from the relevant community. A court may find that an industry standard or custom is set either too high or too low, but being the only actor with poor practices is risky.

4.4 RECOMMENDATIONS FOR ADVERSARIAL ATTACK TOOL MAKERS

Drawing on lessons learned from the Hitman's Quarry, our framework, and three areas of the law, we offer these concrete recommendations for communications best practices for adversarial attack tool makers:

- Help users judge the risk of failure.

[5] *See* Restatement 2d. of Torts, §402A(2).
[6] *See* 3 E-Commerce and Internet Law 27.07 (2020).
[7] 15 U.S.C. §45(a)(1).
[8] Restatement 2d. of Torts, §282.

- Warn users about the unexpected.
- Be clear about present and future limitations.
- Use easy-to-understand language and avoid technical jargon.
- Adopt community best practices.

5 CONCLUSION

Makers of adversarial tools can and should take deliberate action to avoid harming the individuals they seek to protect. Lessons drawn from several areas of the law and from careful analysis of use case scenarios point in the same direction. One of the best things tool makers can do is to give users good, honest, thorough, easy to understand information so that users are well positioned to evaluate the level of protection that the tool provides, and the risk of relying on it, not just today, but over time.

ACKNOWLEDGMENTS

We would like to acknowledge the 2018 Assembly Cohort at The Berkman Klein Center and MIT Media Lab, the 2018 equalAIs team, and Kendra Albert at the Harvard Law School Cyberlaw Clinic for their inspiration and support.

REFERENCES

Dhaval Adjodah, Gretchen Greene, Joshua Joseph, Thomas Miano, Francisco O., and Daniel Pedraza. 2018 MIT/Harvard Assembly Showcase: equalAIs Final Presentation, Apr 2018. URL http://doi.org/10.5281/zenodo.4558005.

Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John P Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=hJmtwocEqzc.

Kristen M. Driskell. Identity confidentiality for women fleeing domestic violence. *20 Hastings Women's L.J. 129*, 2009. URL https://repository.uchastings.edu/hwlj/vol20/iss1/6.

Siddharth Garg. Also adversarial examples as attacks need to work only once to cause bad stuff to happen, adversarial examples as a *defense* need to work *always* (or almost always) to be useful. very different bars. *[Twitter]*, Dec 2020. URL https://twitter.com/sg1753/status/1337242169787211777.

Google Cloud. Cloud Vision API | Google Cloud, 2021. URL https://cloud.google.com/vision/docs/detecting-faces.

Adam Harvey. Computer vision dazzle camouflage. *https://cvdazzle.com//*, 2021. [Accessed 28 February 2021].

Kashmir Hill. The secretive company that might end privacy as we know it. *The New York Times*, Jan 2020. URL https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

Alice Hines. How normal people deployed facial recognition on capitol hill protesters. *Vice*, Feb 2021. URL https://www.vice.com/en/article/4ad5k3/how-normal-people-deployed-facial-recognition-on-capitol-hill-protesters.

Kairos. Face Recognition, 2021. URL https://www.kairos.com/.

Thomas Miano. equalAIs ? Empowering People & Thwarting Machines, Apr 2018. URL https://doi.org/10.5281/zenodo.4554439.

Microsoft Azure. Facial recognition, 2021. URL https://azure.microsoft.com/en-us/services/cognitive-services/face/.

Daniel Pedraza, Dhaval Adjodah, Gretchen Greene, Josh Joseph, Thom Miano, and Francisco O. equalais. *http://equalais.media.mit.edu*, 2018. [Internet Archive: Accessed 3 May 2018 https://web.archive.org/web/20180503002121/http://equalais.media.mit.edu:80/].

Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Fawkes: Protecting privacy against unauthorized deep learning models, 2020. Official Code: https://github.com/Shawn-Shan/fawkes.

Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Image "cloaking" for personal privacy. *https://sandlab.cs.uchicago.edu/fawkes/*, 2021. [Accessed 28 January 2021].

Jane Wakefield. Pimeyes facial recognition website 'could be used by stalkers'. *BBC*, Jun 2020. URL https://www.bbc.com/news/technology-53007510.