Human-AI: Using Threat Intelligence to Expose Deepfakes and the Exploitation of Psychology

Tiffany Mary Tremont

Capitol Technology University

Chair: Dr. Ian McAndrew, FRAeS

February 2, 2023

Human-AI: Using Threat Intelligence to Expose Deepfakes and the Exploitation of Psychology

Approved:

Dr. Ian McAndrew, Ph.D. FRAeS, Chair

Dr. Richard Baker, Ph.D., External Examiner

Accepted and Signed:

Richard E. Baker February 20, 2023

Richard Baker, Ph.D., FRAeS

Date

Ian McAndrew 20 February 2023

Ian McAndrew, Ph.D., FRAeS Date

Dean, Doctoral Programs

Capitol Technology University

ABSTRACT

This qualitative cyber threat deterrence theory research paper outlined countermeasures used to fight against psychological warfare deepfakes using facial recognition software to aid cybersecurity professionals in decision-making within private industry. Using Threat Intelligence to Expose Deepfakes and the Exploitation of Psychology, this research defined the single population and sampling strategy for this suggested qualitative study on Human artificial intelligence (AI). The sample and population of this research paper assisted in determining the breadth of cross-discipline cybersecurity professional subject matter experts (SMEs) sampling required for a comprehensive, non-biased investigation. Only qualified cybersecurity and threat intelligence professionals in the private industry participated in the study conducted entirely online. Participants were prescreened to ensure the target population's validity and improve data quality. A survey was created using an online survey tool called Survey Monkey. This survey was administered to capture participants' responses on how their tools, tactics, techniques, and procedures (TTP), strategies, and methods are used in defense against AI-enabled deepfake malware exploiting GAN for psychological warfare. Research questions answered this investigation's findings, and the contributions confirmed theories. As a result of these findings, we now understand how AI-based technology and countermeasures help private industry cybersecurity, and threat intelligence professionals defend against attacks. Future qualitative research on this topic may incorporate a case study research design methodology.

Keywords: artificial intelligence, advanced persistent threat, cognition, exploitation, decision-making, deepfake, generative adversarial network, human factors, information warfare, malware, metacognition, psychological warfare, threat intelligence, TTP

DEDICATION

With profound gratitude, I dedicate this research to the most cherished people in my life: my devoted husband, Kelly, and my extraordinary sons, Aiden and Liam. The immeasurable love, unwavering encouragement, and unwavering support that you have conferred upon me have been the guiding lights that have led me to the conclusion of this arduous journey and the attainment of my desired doctorate. Your presence has brought tenderness, inspiration, and motivation into my days, propelling me forward despite adversity. You have been my pillar of fortitude every step of the way, providing me with comfort and the confidence that I could overcome any obstacle. This accomplishment is a result of the profound influence your presence has had on my academic pursuits and personal development. Without you, this accomplishment would have remained an impossible dream. Therefore, I offer this humble tribute to you, my cherished family, for without you, this accomplishment would have been impossible.

ACKNOWLEDGMENT

1 Thessalonians 5:16–18 Rejoice always, pray without ceasing, in everything give thanks; for this is the will of God for you in Christ Jesus.

Table of Contents

List of Tables

List of Figures

CHAPTER 1: INTRODUCTION

Artificial Intelligence (AI) malware exploiting facial recognition technology is a serious threat to cybersecurity and society (Greengard, 2020; Jones, 2020; Stoecklin, 2018; Vaccari & Chadwick, 2020). As AI advances, more subtle forms of security risks are included in AI and machine learning (ML), contributing to evasive assaults such as Deeplocker—an AI-powered ultra-targeted malware (Stoecklin, 2018). Cognitive dissonance created using Generative Adversarial Networks (GANs) to magnify deepfakes using photographs, videos, and speech to manipulate thinking damages confidence psychologically (Chesney & Citron, 2019; Jones, 2020; Westerlund, 2019). The multipurpose fabrication of misleading content generated to influence a target audience has a powerful effect on people's beliefs, decisions, and perspectives (Carter et al., 2021). Autoencoders and GANs are examples of neural network designs that produce first-rate copies indistinguishable from the original, making it difficult to tell the real from the fake (Rupareliya, 2020).

Deepfake technology came into the public purview in early 2018 with the infamous video depicting United States President Barack Obama voicing opinions and false statements yet convincingly real by all appearances (Greengard, 2020; Jones, 2020; Vaccari & Chadwick, 2020). There are a variety of reasons to oppose the weaponization of deepfakes. Deepfake makers specifically design counterattacks to disrupt through the volume and velocity of disinformation (Jalaluddin, 2020; Jones, 2020; Peters, 2019). As a result, there are no commonly accepted countermeasure guidelines to assist in defense against psychological warfare deepfakes using facial recognition technologies to aid cybersecurity specialists' decision-making within private enterprise (Aiyanyo et al., 2020; Boivin, 2018; Jones, 2020; Westerlund, 2019).

Background of the Study

Malicious actors have made a name for themselves by using malware, threat vectors, and deepfakes to access private data and systems. Particularly worrisome is the possibility that AI could be used as an attack tool or even as an attack surface to help an adversary launch bigger, more autonomous attacks (Liu and Murphy, 2020). Facial recognition software revolutionizes the verification and interaction with mobile phones, computers, social networking, and other biometrics-required applications (apps). Cybersecurity frameworks and standards help firms identify unlawful assaults and patch newly discovered security flaws. As the digital transition proceeds in society, the convincing authenticity of deepfakes will remain a constant. It is common nowadays to underestimate AI technology's complexity; nevertheless, the breadth and scope of technology used to create deepfakes is the most remarkable aspect of AI's effect (Westerlund, 2020; Fletcher, 2018; Anderson, 2018). Therefore, this study discusses the countermeasures and methods used to defend against AI-based malware injection attacks for psychological weaponization using facial recognition software leading to targeted attacks of AI-enabled malware exploiting in the cybersecurity field. Further, through data gathered from cross-functional cybersecurity experts' implications of AI and its impact on the cybersecurity field in the current landscape: AI ethics, understanding deepfakes, benefits of deepfake and facial recognition technologies, who creates deepfakes, threats deepfakes pose, examples of deepfakes, AI legislation and regulation, policies and proactive behaviors, human performance, and methods used to combat deepfakes and facial recognition technologies.

AI Attack Tool

For destructive targeted assaults, AI, specifically deepfakes as an attack tool, is a source of worry (Langa, 2021).

Liu and Murphy, 2020, outline the benefits of applying AI methods to increase malware detection efficiency. To safeguard against evasive threats, an emphasis on the need for AI awareness raises the concern that exploitation of attack profiles and vulnerabilities may exist. This research will look at the advanced tactics and techniques used in AI malware to protect against psychological warfare deepfakes using facial recognition software to equip cybersecurity experts to make better decisions in the commercial sector. The emphasis is the need to synthesize the data to raise awareness of open-source AI tools utilized in defensive cybersecurity tactics, tools, procedures, approaches, tools, strategies, and procedures.

Deepfake Videos

"Deepfake videos are a present-day concern that will only become more challenging as ML technologies develop, increasing the authentic feel of the videos and making detection more difficult" (Peters, 2019, p.10). Face recognition software might produce volume and velocity in hybrid warfare as a contributor to AI-based malware assaults, as deepfake technology offers risks to malign and develop mistrust. (Jones, 2020). The implications of AI-generated deepfake audio and video technologies were investigated in this study (Jones, 2020). Jones (2020) claims that GAN tools allow users to alter video and audio recordings, making it difficult to determine what is a fake, as supported by Jones (2020).

Problem Statement

The research problem is to determine ways to defend against AI-based malware injection attacks for psychological weaponization using facial recognition software leading to targeted attacks (Jalaluddin, 2020; Jones, 2020).

General Problem

The general problem is that generative adversarial network (GAN) is weaponized through

AI malware to exploit facial recognition software (Boivin, 2018; Greengard, 2020; Jalaluddin, 2020; Jones, 2020; Stoecklin, 2018).

Specific Problem

The specific problem is that there are no widely adopted countermeasures identified against AI-enabled malware exploiting GAN for psychological warfare (Greengard, 2020; Jalaluddin, 2020; Jones, 2020; Labrien, 2016). Consequently, it necessitates the development and implementation of additional protective countermeasures.

Purpose of the Study

The purpose of this qualitative survey research design study is to discover the countermeasures used to defend against psychological warfare deepfakes utilizing facial recognition software to assist cybersecurity professionals in decision-making within private industry (Aiyanyo et al., 2020; Boivin, 2018; Jones, 2020; Westerlund, 2019).

Significance of the Study

Theoretical Significance

The theoretical significance of this study will be a guideline to an established single, flexible framework that is based on the seven-step National Institutes of Standards and Technology (NIST) 800-37 Rev. 2 | Risk Management Framework (RMF) for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy, cyber deterrence and decision-making theory, allowing for rigorous continual mitigation countermeasures against AI-enabled malware exploiting GAN for psychological warfare. This qualitative study's findings are important for providing a future defensive countermeasure framework, assisting cross-functional cybersecurity subject matter experts (SMEs), communities, and societies in identifying, assessing, and prioritizing risk and residual risk, and forming a coordinated and economical application of resources to minimize, monitor, and control the probability and impact of adverse events.

Practical Significance

The practical significance is that if the resulting guideline is implemented, it will holistically benefit the community and society as a whole by developing a new defensive countermeasures guideline to supplement the RMF and serve as a foundation for cybersecurity subject matter expert (SME) decision-making. However, with the growth in cyber capabilities of both state and non-state actors, there is a great urgency to develop an approach that is not only scientifically sound but also operationally feasible. The analytical ML technique of utilizing GANs to create impossibly realistic visuals avoids the issue of facial recognition software. Future generations will be able to use this guideline as a reference, with or without extensive interpretation. Face recognition software used to unlock smartphones and computers, raises the possibility that AI-enabled malware will leverage deepfake technology by replacing a person's face and voice, or a combination of both, as Seker (2020) suggests. Despite this progress, the future of AI science is far from certain. Understanding how GANs generate deepfakes benefits cybersecurity professionals and specialists in detecting and defending against them. Future cybersecurity professionals may use this as a reference to defend against AI-enabled malware, preventing the exploitation of AI deepfake technology.

Theoretical and Practical Significance Relationship

The theoretical significance stems primarily from discovering a solution set of open-source countermeasures for protection and defense against the malicious use of AI facial recognition technologies. Furthermore, the practical significance of the findings affects how GAN-created deepfakes might benefit cybersecurity experts in defense and detection.

Specifically, the theoretical significance will be modelled after the NIST 800-37 Rev. 2 RMF to allow a methodical approach, as reflected by the respondent data and countermeasures provided used in response to AI-enabled malware, threat intelligence, and future deepfakes leveraging psychological cognition and perception bias. Controls are not repairs; rather, controls are automated or manual processes intended to avoid, identify, or remedy malicious activity. The elimination of tangible and intangible inconsistencies, as well as the search for disparate technologies, duplicate controls, or other inefficiencies in existing controls, is conducive to good outcomes in risk mitigation strategies due to the relationship between theoretical and practical significance. The risk treatment outlined in the study's guideline must be proportionate to the estimated target's worth. Adopting the guideline will define and assign control attributes, allowing for the deployment of effective and efficient control schemas that will thwart emerging threats.

Psychological Warfare Deepfakes

It is crucial to recognize and defend against psychological warfare deepfakes utilizing facial recognition software. Deepfake psychological warfare is an emerging weapon that has the potential to wreak havoc on a mass scale. GANS are a biometric standard recognized in the applicable International Organization of Standards (ISO) 30107. Standards provide baseline definitions for identity attributes of a living being such as fingerprints, iris, voice, and face. Seker (2020) agrees that detection methods to identify or circumvent verification procedures to detect deepfakes are a framework suitable for biometric software applications' security.

Since the inception of deepfake, researchers are working on a new framework for techniques to detect GANS by establishing statistical distribution characteristics (Seker, 2020). Software applications need to be made secure against malicious attacks**.** This research is unique