Generative Artificial Intelligence and Its Relationship with Disinformation

By

Andrew Aurand

College of Technology, Wilmington University

August 2023

Submitted In Partial Fulfillment of the Requirements for the Degree of Master of Science in Cybersecurity

Table of Contents

List of Illustrative Materials

ACKNOWLEDGEMENTS

This acknowledgement section is in no particular order.

Firstly, I want to thank my wife, Abigail Follis. Without her, I never would have never started my bachelors, nor would I have continued with higher education to get my master's degree.

Secondly, I would like to thank Dr. Mark Hufe for running an amazing Python course and having faith in me as an adjunct professor.

Thirdly, I would like to thank Dr. Jim Fraley for extending the same faith as Dr. Hufe with respect to my instruction, as well as running a special topics course that opened my eyes to disinformation/misinformation.

Next, I would like to thank Professor Cody Dostal and Professor Fred Stinchcombe for running outstanding courses at Wilmington University and making my transition from undergraduate to graduate a smooth one.

Lastly, I would like to personally thank the companies responsible for generative AI. Without these companies' disregard for ethics in chasing the almighty dollar, I would not have been able to write this paper.

Abstract

As generative artificial intelligence (AI) has become available to the public, it has wrought significant problems in nearly every sector of society. Among security professionals, it has become an issue in its usage within disinformation campaigns and attacks. This research hypothesizes that increased reporting on AI generative software has resulted in increased social engineering attacks and mass disinformation events. It argues that the ability to mask identity and maintain anonymity reduces the risk of the attack, leading to malicious actors adapting AI projects to their needs. Through a series of case studies, this study concludes that the perception of AI and its unmitigated use has increased its use in disinformation attacks. This study recommends increased education, government oversight, and ethical standards for all present and future AI projects, thus mitigating its use as a tool for spreading disinformation and misinformation.

Keywords: cybersecurity, generative artificial intelligence, disinformation, misinformation, social engineering

INTRODUCTION

In 2023, the pictures as shown in appendix A were circulated through popular social media sites. They depict Pope Francis wearing a puffy coat with a decorative crucifix hanging from a chain. The clothes are modern, not something expected of one of the highest religious authorities in the world. Deeply pious Catholics need not fret, though – the image was created by a generative artificial intelligence program (AI) called MidJourney (Huang, 2023).

Although a trained eye could pick out some inconsistencies or things that do not appear quite right in the image, at first glance it looks disturbingly real. In recent years, AI generative technology has exploded on the market, used to generate images, art, and text. Its capabilities seem limitless, as students use it to write their essays or programmers ask it for assistance in revising code. The abilities of this software have generated controversy and danger. This study will examine the ways in which AI generative software is being used by malicious actors to create harmful disinformation and why it should be taken seriously by cybersecurity professionals. The findings and argument presented regarding the relationship between social engineering, disinformation, and generative AI should be taken seriously by security professionals working with public-facing organizations to protect their infrastructure and the safety of society in general.

Although ridiculous images or jokes related to AI may make the rounds on the Internet, unmonitored and unmitigated public use of AI presents a real and present danger to both public and private organizations (Chu et al., 2020; Turchin, 2019; Yudowsky, 2008). Malicious actors are able to mimic the verbal and written communication of individuals or organizations, gaining access to vital information or demanding ransom in fabricated crimes. These actions waste public

resources and cost organizations thousands.[1] Since these attacks are a form of social engineering utilizing established computer networks and cyber connections, it is imperative that cybersecurity professionals take the time to create incident response plans to AI generated threats. Additionally, private and public organizations need to acknowledge the clear and present danger, which must be mitigated by developing a clearer understanding of its capabilities.

Generative AI

Understanding the history of generative AI is vital to grasping its rapid development. It was first introduced in the form of a chatbot in the 1960's. This first chatbot was named ELIZA, developed by MIT researcher Joseph Weizenbaum at their Artificial Intelligence laboratory.[2] The intention was to create an AI chatbot that simulated human conversation. Although rudimentary by today's standards, it set the stage for what we know today as generative artificial intelligence. Wiezenbaum's ELIZA was programmed to behave as a psychotherapist, asking the user to elaborate on their feelings. ELIZA would learn new words or phrases from the users, frequently asking for elaboration in a manner reminiscent of a therapist. Wiezenbaum was surprised at the way people reacted to the chatbot, particularly in how they treated it as if they were speaking to a real person. This early example of how a machine can mimic human interaction enough to cause non-technical professionals to react as they did demonstrates the capabilities of such systems. It exemplifies how social engineers can use these technologies to manipulate organizations or people.

Most of the artificial intelligence research from the 1960's to 1993 was kept in academia since the Internet was not available to the public. From 1993 onward, with the release of the

---

[1] See Appendix B for more detailed reports on the costs associated with pursuing these actors and related criminal activity.

[2] ELIZA is named after the character Eliza Doolittle in the Irish play *Pygmalion*. Despite its presentation in all capital letters, it is not an acronym.

world wide web and interconnective world, we have seen an explosion of various AI chatbots. One most Internet natives would recognize is Cleverbot, released to the public in 2008. Mimicking user speech, it could hold loose (albeit often stilted or strange) conversations with users. Similar to ELIZA, it learned new phrases from its userbase, mimicking human interactions. At its height, it became so effective that users questioned whether they were "speaking" to a chatbot or paired with another user. This led to a rash of instances where users attempted to convince Cleverbot it was a chatbot (Gehl, 2014). As users became more familiar with Cleverbot and its capabilities, it became clear that a machine could interact with the public on the Internet in a way that caused people to doubt its existence as a machine.

2014 was a turning point with the introduction of generative adversarial networks (GANs). GANs are a form of machine learning algorithms. They allows a computer scientist to take large amounts of data (photos, videos) and create a dataset to "train" the artificial intelligence program. Since then, large language models (LLMs) have been in the spotlight, allowing large amounts of text to be used in datasets, which can be used to synthesize results based on a prompt (Fruhlinger, 2023). At the time of writing, several generative AI programs exist, and allow a user to accomplish various tasks. These programs include ChatGPT, ElevenLabs, GitHub Copilot, and MidJourney. Each one has a particular skillset, ranging from generating text to creating images, but they can be used in conjunction with one another. As users continue to use them or information is added to the Internet, the software "learns" better ways to communicate and mimic human expression and experiences. Details on each of these programs will be provided in the following section.

Social Engineering/Disinformation

Social engineering is the act of using manipulation to make users act a certain way or obtain certain goals. Some social engineering can be considered "good," such as targeted campaigns to get children to not smoke or do drugs. Other social engineering attempts are malicious, such as tricking a user into divulging sensitive information (*What Is Social Engineering - The Human Element in the Technology Scam| Cybersecurity | CompTIA*, n.d.). Social engineering can take many forms, such as spam, phishing, whaling, and voice phishing (Vishing). The medium changes, but the end goal remains the same: Manipulating an individual for a specific purpose, usually capital gain.

A true historical look at social engineering would easily look back hundreds of years in and span several books. For the purposes of this study, I will focus on social engineering through the lens of a cyber security professional, beginning from the 1990's onward. Social engineering in the age of early computing was one of the easiest ways to gain information about a system, computer, or network. Kevin Mitnick is one of the more famous (or infamous) social engineers who thrust a spotlight on the niche field. Using social engineering, Mitnick was able to gain access to passwords and privileged information from various companies and the government. He stole source code from several notable companies in the mid 90's, such as Sun Microsystems, Nokia, and Motorola Corporation just using a phone and social engineering techniques (*Who Are Hackers - The Testimony of An Ex-Hacker | Hackers | FRONTLINE | PBS*, n.d.). In doing so, he ended up serving prison time and eventually testified before congress during a hearing detailing federal information systems in March of 2000. Utilizing the human element of security, he was able to do as much if not more damage than a computer virus or trojan. Mitnick claimed the

weakest link in computer security is not the computer, nor the code it runs from, but the human element.

Traditionally, social engineering has been interpreted as actions taken to garner information to instigate an attack on an organization using the human element of an organization. Actions that do not involve using a coded vulnerability or brute-forcing a password are generally classified as social engineering. Academic literature on the subject has hesitated to relate the spread of disinformation to social engineering, since it is not always done to gain access for profit to a particular organization. Recent activity in the private sector involving the falsification of information to demand a ransom or other financial benefit demonstrates that spreading disinformation should be considered a type of social engineering.

Mass disinformation is a form of social engineering. Mass disinformation could be malicious actors spreading a fake video, doctored photo, or wrong information while using an algorithm to get as many eyes in front of the media as possible. It was not traditionally the domain of cybersecurity professionals, but in recent years has slowly dominated the conversation. Disinformation differs from misinformation in its intent; disinformation exists to intentionally mislead its audience, while misinformation is usually spread by parties unaware of its lies.

Generative artificial intelligence and social engineering/disinformation are intrinsically linked. Malicious actors use cutting edge technology and techniques to accomplish their goals. Having generative AI in their toolbox allows them to quickly and acutely spread disinformation and attempt social engineering attacks on a scale the security industry has not seen previously. According to the cybersecurity company KnowBe4, social engineering attacks utilizing

generative AI increased by 135% since the widespread availability of various AI platforms (Sjouwerman, n.d.).

Research Question

Public use of AI has generated numerous controversies as it grows in its capabilities. To better address the problems it creates, they must be separated and broken down into their components. In the case of the spread of disinformation, there exists abundant literature on its effects in different sectors of society. Social engineering's relationship with disinformation has also received due attention. Academic study on AI generative software, however, is in its infancy, beyond some early speculative research from the twentieth century. Thus, there is a gap in research that examines how AI generative software has affected the prevalence of disinformation attacks. This leads to the central question of this thesis:

> *How has AI generative software impacted the frequency and ferocity of disinformation attacks?*

In April 2023, a woman named Jennifer DeStefano received a call that every parent dreads. The call came from her daughter, claiming she had been kidnapped. A man's voice accompanied her daughter's, panicking DeStefano. The man in the call began demanding a ransom for the return of her daughter and threatening violence if DeStefano contacted law enforcement. Panicked, DeStefano immediately contacted authorities, who quickly realized it was a scam after reaching the daughter, who was safe. Authorities and DeStefano learned that a malicious actor had cloned her daughter's voice using information from social media, hoping to scam her out of $1 million (Karimi, 2023).

This example of cruel social engineering and manipulation demonstrates the capabilities of AI in the hands of a savvy malicious actor. Thankfully, DeStefano's daughter was safe, but it

shows how the information on the Internet is being abused to exploit children and their families. Additionally, the calls and law enforcement deployment cost the public in tax revenue, all ultimately a waste. The non-monetary impact is an increased level of fear and trauma within the public, which has been proven to lead to less safe or productive communities. Therefore, the capabilities of generative AI and its uses when spreading disinformation should be taken seriously by both law enforcement and security professionals. Developing an effective security plan will require the cooperation of different groups and the resources available to industry specialists.

AI generative software, due to its widespread availability, has led to a massive increase in disinformation, misinformation, and attempted social engineering attacks. This became a problem due to unfettered access to these AI programs. Private companies released various platforms (ChatGPT, MidJourney, Elevenlabs) with no safety features, no oversight, and initially free of charge. Unfettered capitalism took over, and individuals immediately started devising ways to make money. Corporations, on the other hand, started laying off writing staff, claiming ChatGPT will make them obsolete (Kumar, 2023). Malicious actors started using various generative AI programs to spread disinformation, scam users, and sow chaos on the Internet connected world. They were given unbound access to these programs due to the lack of safeguards implemented when designing these systems. In some cases, these programs created problems that only affect small groups, such as teachers facing rampant cheating as students generate their essays through ChatGPT. The problems have since increased. In one case, a security researcher used an AI generated voice of himself to access his bank account, since his bank used voiceprint identification. He theorized that with widespread availability of AI

generated voices, malicious actors can easily mimic a person and steal their money or personal information (Cox, 2023).

Problems surrounding generative AI exist in almost every industry. Below is a table outlining the industry, and potential problems with AI software:

Industry	Negative Effects of AI software
Education	Plagiarism, Cheating on written essays
Writers (book, website, tv shows)	Job cuts, use of AI to write entire stories, TV shows
Artists (Traditional Pictures, photography)	Create derivative works without credit. Some corporations have fired artists, since using AI is cheaper than an employee
Computer Science (Programmers)	Companies laying off programmers, Junior programmers turning to ChatGPT/GitHub Copilot to do less work
Voice actors (animated shows/commercials	Companies claim voice actors' likeness (voice) for use in perpetuity. Usage of ElevenLabs to clone voice actor's voice, then not compensate them
General Public	Distrust in the overall technology field due to rampant disinformation, stolen works, and bad press

Table 1 Broad examination of the effect of generative AI on varying industries. Information pulled from news bulletins and reports.

Per Table 1, Generative AI is affecting large swathes of the world in various industries. Several deficiencies exist in current academic research over the ethical and cybersecurity implications in AI. Most AI generative programs were not produced in an academic setting, but in the private sector. The wider academic community was given access to these programs around the same time as the public, much to the dismay of the cybersecurity academic community. At the time of writing, all of the AI programs are closed off to scrutiny by computer scientists and cybersecurity professionals. The datasets used to train these AI platforms are considered confidential information, and any safeguards implemented into these programs are also considered confidential, so no auditing can be done.

Since academics are refused entry for study, literature covering recent developments in AI has been sparse in those spaces. Industry journalists, though, have produced some analysis that could prove fruitful to cybersecurity researchers. Through these resources, we can form a more cohesive picture of the effects of generative AI on the spread of disinformation.

Increased reporting on and access to AI generative software has resulted in increased social engineering attacks and mass disinformation events. The ability to mask identity and maintain anonymity reduces the purported risk of an attack, leading to malicious actors adapting AI projects to their own needs. I will be analyzing several cases where it is believed generative AI was used to spread disinformation and social engineering attacks. This hypothesis looks at one small slice of how generative AI is used, to reach conclusions on how to best thwart disinformation and social engineering attacks on individuals and organizations. This paper will look at academic papers, industry experts, and original research in the form of case studies.

LITERATURE REVIEW

The literature for this study will cover three major themes: disinformation, generative artificial intelligence (AI), and social engineering. These three themes connect in their relationship with multidisciplinary approaches, particularly among the social science. It is clear that studies on disinformation and social engineering relate, but the recency of generative AI has limited the ways in which it can be studied or tied to other areas of research. Since current academic literature regarding AI is in its infancy, this research will examine case studies that provide further detail on its capabilities and its reception in academia and industry. Therefore, literature that exemplifies the advantages and disadvantages will also be examined.

Disinformation

Edward Louis Bernays was an American academic, considered in some circles the father of public relations. For others, he is known as the father of propaganda, mass manipulation, and mass marketing. Bernays utilized social engineering techniques to advertise different products, manipulating consumers into believing a good or service would better their lives to an extreme degree. For example, in 1929, Bernays launched a campaign that told suffragettes that smoking tobacco empowered them and put them on an equal playing field with the men of society. His campaign took hold among women seeking greater agency and opened a new market for tobacco companies. While most of Bernays' work focused on developing catchy advertising campaigns, he looked specifically to manipulate a sector of the public for marketing purposes. His tactics are reminiscent of mass disinformation campaigns, which are often done in the interest of profit. Bernays passed in 1995, but his legacy has permeated American marketing culture and is an essential aspect to understanding the development of mass disinformation.

Although misinformation and disinformation are closely related, they differ in their intent. Both spread inaccurate information that could be interpreted as fact, usually harmful to the individuals consuming it. Misinformation is incorrect information that is spread unintentionally, usually through sharing on social media or via word of mouth in offline spaces. For example, during the tragedy of the Boston Marathon Bombing, the social media site Reddit determined they had found the culprits during the manhunt. Their assurance was based on nothing more than appearance, but the "fact" of the identity of the perpetuators spread throughout the website. Soon, users were directly contacting law enforcement in Boston to inform them they had found the culprit. These actions were not done in malice towards the soon-to-be-found innocent man Reddit users had identified. Rather, they were done with the intention of catching the offenders but were based upon misinformation. The man accused of being the bomber suffered from online and offline harassment for years, demonstrating the harmful effects of misinformation (Myles et al, 2018).

Disinformation also describes the spread of inaccurate information but is done so with the intention of spreading a lie. Individuals that generate this information intend for others to be deceived to achieve their own ends. For example, if a grade school bully spreads a rumor about a new student, that is a case of disinformation. The bully has no evidence to suggest the rumor about the student is true, but does so with the intention of causing harm, gaining attention, or creating a common enemy that they can rally against. Although this example is relatively benign, cases of disinformation have led to financial and bodily harm to people in the United States. It is thus important to take instances of disinformation seriously.

Although disinformation is related to cybersecurity, it is also understood through the lens of different disciplines. Cybersecurity and psychology, for instance, have become increasingly

intertwined in recent years. Called Cognition Security (CogSec), this newly developed field examines the ways fake news and disinformation impact human cognition (B. Guo et al., 2020). Examining disinformation from a multidisciplinary approach requires cybersecurity professionals to communicate with other fields to deepen their understanding of the dangers of disinformation. Understanding the theories around motivations of users can assist security professionals in developing a more holistic picture of the effects of disinformation. Additionally, it can provide a roadmap to explore its origins and prevent it from causing harm.

Researchers face multiple challenges when attempting to develop plans to protect the CogSec of the general public. These challenges exist in four categories: human-content cognition mechanisms, social influence and opinion diffusion, fake news detection, and malicious bot detection (B. Guo et al., 2020). These elements are closely related to exploring the psychological and sociological elements of social engineering, exemplifying that disinformation has a place within the subfield. Future research must come from the combined efforts of researchers across disciplines, including those in AI research.

As social media and Internet-connected devices become ubiquitous, so does disinformation. Security researchers have become increasingly concerned with various forms of social media disinformation, how they are created, and efforts to combat them (Shu et al, 2020). Since disinformation is so varied, it must be analyzed in a way that appreciates those differences. Generative Adversarial Network (GAN) generated fake images, fake videos and "deepfakes,"[3] and multimodal content are a few examples of the types of disinformation. The vast majority of these types of disinformation is spread through the Internet due to the low or no cost of

[3] A deepfake is an image or video – more commonly the latter – in which a person or object has been placed in a scenario in which they were not present previously. For example, utilizing the wealth of videos of a particular celebrity, an individual could create a video that shows that person flying an airplane, despite the celebrity having never taken flying lessons.

publication. “Disinformation farms” easily create thousands of fake accounts (also called bot accounts), generate a GAN profile picture, and spread inaccurate information via major social media sites (Shu et al., 2020).

Although various methods exist to detect disinformation, such as modeling user interactions, using user sentiments, and leveraging the content to detection information, most social media sites use one method instead of a multi-faceted approach (Shu et al, 2020). The problem is complex and nearly impossible to fully mitigate, especially regarding fake news. Interdisciplinary and cooperative efforts are encouraged to create solutions. As security researchers learn more about the dangers of disinformation and misinformation, it is imperative that professionals take the threat more seriously (Shu et al., 2020; B. Guo et al., 2020; Andrews, 2021).

The social media age has contributed significantly to the prevalence of disinformation, particularly as it relates to AI generated images. As companies seek profits by increased engagement (and, therefore, advertising money), they push information that would appeal to users. If users feel a connection to the content they are viewing, they are more likely to remain on the site and stay engaged (Sharma et al, 2020). Algorithms inadvertently create “filter bubbles or echo chambers” of disinformation, where communities of users engage in the spread of misinformation (Shu et al., 2020). The existence of these spaces makes it difficult to find the source of disinformation, since users are often convinced of the truth of them within their own online spaces.

The turbulent socio-political environment of the early 2020s brought renewed interest to “fake news,” the most well-known form of disinformation. Disinformation of this caliber manipulates viewers to believe a particular account that could be harmful to their wellbeing

(Andrews, 2021). The researcher Crispin Andrews identified the development of "engineering consent," in which marketers manipulate a viewer to buy into a new product or habit that would prove profitable to an organization but detrimental to the consumer's health. For instance, Andrews examines the way in which Bernays used the psychological practices of Sigmund Freud to manipulate women into smoking tobacco products.

With social media growing as a center for the spread of disinformation, studies regarding its spread through particular networks have risen (B. Guo et al, 2020; Shu et al, 2020; Z. Guo et al, 2022). Zhen Guo et al (2022) used a game-theoretic model to investigate how individuals process information and how that impacts the dissemination of disinformation. They conclude that "uncertainty-based [opinion models (OMs)] may assist users in excluding uncertain information and believe true information" (Z. Guo et al, 2022). Propagation of disinformation may cause the decrease of social capital, which impacts a person's place in a community and their perception of themselves.

Analysis of subjective interactions between users or with users and a general community is virtually impossible to objectively interpret. Therefore, researchers have turned to models developed by other social scientists (such as game theory) to find commonalities in online interaction. Z. Guo et al (2022) utilized this model in a way that assumes two opposing viewpoints, one that is riding on disinformation and another that has a more factual argument. This experiment – while useful to quantify the experiences of people within an online space – neglects the presence of echo chambers or the possibility of other players within a particular system. The complexity of these networks makes it difficult to detect disinformation, let alone determine the players in a provided space (Shu et al, 2020).

Building on the research conducted by Z. Guo et al. (2022), Carlos Diaz Ruiz and Tomas Nilsson demonstrate how disinformation can generate through echo chambers in social media networks. They propose a two-phase framework that shows how disinformation spreads. The first phase is "seeding," where a malicious actor inserts deceptions by masquerading their legitimacy. For example, this actor could pretend to be someone with legitimate medical credentials while discussing the dangers of a life-saving medication. The second phase is "echoing," where the malicious actor convinces participants of the content to repeat the information, spreading their disinformation through an organized campaign. As users spread this information, it moves from disinformation to misinformation, demonstrating the relationship between the two concepts. Diaz Ruiz and Nilsson discovered that methods of spreading disinformation worked better than others, such as identity-driven controversies being more effective than those associated with less personally relatable content. Utilizing rhetorical theory, we learn that arguments drive a model that can identify valid knowledge within an echo chamber (Diaz Ruiz & Nilsson, 2018).

Narrowing down the definition of disinformation is central to forming an adequate analysis of its relationship with social engineering. Disinformation encapsulates false information that is spread deliberately with the intention to push a particular narrative for profit. Disinformation can therefore come in a variety of different forms, some of which blur the line between disinformation and misinformation. Conspiracy theories, for example, typically involve someone rejecting a common interpretation of an event, attributing it instead to a malicious, secret society (Sunstein & Vermeule, 2009). These often run parallel to rumors, although rumors typically do not have a conspiratorial element to them. Additionally, rumors tend to involve smaller communities or single individuals rather than events.

These two aspects of disinformation – rumors and conspiracy theories – have led researchers to their own paths of analysis. The complexity of these two terms is beyond the bounds of this study, but the less debated type of disinformation – fake news – will be covered more thoroughly. Fake news is defined as a deliberate attempt to mislead readers by making up facts. The key element, making it almost "interchangeable" with disinformation, is in its intent to misinform readers (Shu et al, 2020). This analysis will not use fake news and disinformation similarly but clarify when a piece of AI generated content refers to fake news or not.

Social Engineering

Social engineering is a broad and frequently fluctuating subfield of cybersecurity. Among professionals, it is often used in contexts where organizations attempt to prevent phishing or employees mistakenly sending money to malicious actors. Research in social engineering stretches across multiple disciplines, namely sociology and psychology. Cybersecurity research tends to examine it from a practical perspective rather than a deep dive into human interaction and consciousness. For the purposes of this study, the perception of social engineering will be explored across major contributions to the literature within cybersecurity.

Since there is a debate about what constitutes social engineering among professionals, it is necessary to explore how the subfield is treated among academics and industry professionals. This research primarily examines the effects of unmitigated access to generative AI on private organizations and individuals, which I argue is related to studies on social engineering. This argument hinges on the spread of disinformation acting as a form of social engineering, thus a brief exploration of the commonly referenced literature on social engineering is necessary.

Kevin Mitnick is regarded as one of the forerunners of social engineering in cybersecurity. As prolific as he is controversial, any analysis of social engineering would be

remiss to not acknowledge his contributions. In his largely fictional book, Mitnick details several plausible social engineering scenarios, how they would work, and the best ways to defeat them. Published in 2002, the details therein are dated and do not address widespread Internet use. However, it did spark conversations within the budding cybersecurity community. It acknowledged that social engineering was a real threat in cybersecurity and presented convincing evidence to that fact. Security professionals had to acknowledge that interpersonal skills and education would be a key factor in educating organizations, governments, and individuals to maintain secure private or proprietary information.

Although largely fiction, Mitnick's methodologies discussed are certainly possible. The scenarios he outlines are recorded as having been used throughout the 2000s and 2010s by malicious actors and ethical white hats in their attempts to social engineer users. In May 2021, a con-artist in the UK used vishing (voice phishing) to fool people into signing up for a vaccine program. The program was not real, and the perpetrator used the information gathered from victims to wipe their bank accounts. This type of attack is similar to the "Let Me Help You" attacks described by Mitnick. Although Mitnick did not provide real-world examples in his work, there are instances – such as the fake vaccine program – where his strategies were implemented.

As social engineering has developed alongside computer processing technology, machines have become increasingly intertwined with social engineering tactics. Aroyo et al provides insight into how trust towards inanimate objects could be used to divulge personal information, which can then be used to manipulate victims. In this specific study, researchers used a humanoid robot (named iCub) and Kevin Mitnick's social engineering framework to attempt the following: collect personal information, develop trust and rapport with the participant, and attempt to exploit the gained trust of the participant (Aroyo et al, 2018). In the

scenario, iCub would help a user with a treasure hunt, while getting them to divulge sensitive information. Upon finding the treasure, iCub would attempt to convince the participant to gamble the treasure they won, leveraging previously disclosed personal information. The results of this study showed that people tended to build rapport and trust with iCub and gambled without question when asked to.

The study conducted by Aroyo et al demonstrates the trust that users are willing to give to something they know is a machine. Sociological studies have shown that the human mind does not always differentiate between a "robot" and an actual user (Johnson, 1988; Prasad, 1994; Borch & Min, 2022). Research has tended to examine human-machine relationships in various settings, especially as it relates to best business practices. The implicit desire to trust causes users to confide information that would otherwise be sensitive, which can then be exploited, such as in the case of human interaction with chatbots.

Interpretations of social engineering by cybersecurity professionals typically involve manuals to understand its nuance. Christopher Hadnagy published the common tactics, techniques, and procedures (TTPs) in his book in 2018 that contemporary cybersecurity professionals commonly use. Part manual and part psychology book, the TTPs are standard practice by most social engineers today. The psychology-centered aspects of the book demonstrate the relationship between cybersecurity and the social sciences and the importance of a multidisciplinary approach.

Joe Gray published *Practical Social Engineering* in 2022, providing a blueprint for the ways an ethical hacker can socially engineer an organization or individual. The book is primarily used by penetration testers to conduct an appropriate security audit. It also analyzes the ethical concerns surrounding social engineering. Gray argues that ethical social engineers can become so

by following local and national laws, paying particular attention to how they differ based on the target's location (Gray, 2022). Additionally, he recommends openness with the target after the social engineering investigation to avoid miscommunication.

Although not strictly academic sources, the books authored by Gray, Hadnagy, and Mitnick demonstrate the most common interpretations of social engineering for cybersecurity professionals. Each book attempts to be the most comprehensive in its interpretation of social engineering. What each is missing, though, is the inclusion of disinformation. Since disinformation has a clear intent to spread false information and cause harm through it, it should be considered an aspect of social engineering. Some professionals have shied away from adopting this interpretation, largely because it further complicates social engineering. This analysis will argue that disinformation belongs under the umbrella of social engineering and should be taken seriously among cybersecurity professionals that analyze the two subjects.

Generative Artificial Intelligence

Literature on generative AI is growing at a rapid rate. Some studies from attempts at AI prior to the Information Age do exist, but they mostly explore the probability of true artificial intelligence or speculate on the state of STEM fields (Smithers, 1988; Göranzon et al, 1988; Nuki, 1990). Following the introduction of ChatGPT in November 2018, generative AI exploded into the public consciousness and, subsequently, among security researchers. Due to its relative newness, though, there is little in the way of substantial academic literature. Contrary to popular academic opinions, though, industry literature often has the jump and does not lag behind (Hagendoff & Meding, 2021). Thus, current research has several gaps and is often vague, laying the foundation for more in-depth research.

Irene Solaiman discusses a useful tiered framework allowed for researcher access to generative systems. She proposes a framework with six levels of access to generative AI systems: fully closed; gradual or staged access; hosted access; cloud-based or API access; downloadable access; and fully open" (Solaiman, 2023). Within these gradients, several tradeoffs are made such as the availability of generative AI to the public and independent researchers.

Figure 1 in Appendix B is a replica of Solaiman's gradient chart, demonstrating the levels of access for generative AI software. It exemplifies how, as systems become more open, they better enable audits and community-based research but become more difficult to control. Similar to problems facing disinformation, the research conducted by Solaiman demonstrates the necessity of working with "multidisciplinary experts and the AI community" (Solaiman, 2023). Thus, studies on generative AI must approach it from a multidisciplinary perspective, otherwise questions will not receive their appropriate answers or speculations.

Adopting the Dual Use of Research Concern (DURC) framework can provide a deeper understanding on social awareness of generative AI (Grinbaum and Adomaitis, 2023). DURC is more commonly used in social sciences as a tool to encourage ethics and could apply to Large Language Models (LLMs). Academic conversations around DURC reveal two conflicting elements of the study of security: the pursuit of knowledge and the safeguarding of public safety (Grinbaum and Adomaitis, 2023). As security professionals, we must attempt to address both appropriately. This study will examine the ways that regulating or providing greater researcher access to generative AI can provide more safeguards for the public. Thus, the pursuit of knowledge goes hand in hand with achieving public safety.

Large generative AI models (LGAIMS) are transforming the way we express ideas and communicate (Hacker et al, 2023). In the European Union (EU), there is a lack of regulation,

guardrails, and ethics involved in LGAIMs.[4] Similar to the approach outlined by Solaiman, it is suggested by Hacker et al that a tiered approach be adopted for the regulation of LGAIMs. This approach would cover four areas: direct regulation, data protection, content moderation, and policy proposals. These areas are supported by "three layers of obligations: minimum standards…high risk obligations….and collaborations along the AI value chain" (Hacker et al, 2023). Due to the lack of laws surrounding specifically LGAIMs, current EU and US regulations do not apply to them and will need to be updated and revised. Going forward, LGAIMs users should be grouped into different categories: LGAIM developers, professional and non-professional users, deployers, and recipients of LGAIM outputs.

Among researchers of AI generative software, there exist two camps: those that believe AI can be used as a public good and those that distrust its development. For AI to function as a public good, it must follow necessary requirements to become "socially good" (Züger & Asghari, 2023). These elements lead academics to argue that AI should be developed with the interests of the people. This "public interest AI" framework consists of five elements: "(1) public justification for the AI systems, (2) an emphasis on equality, (3) deliberation/co-design processes, (4) technical safeguards, and (5) openness to validation" (Züger & Asghari, 2023). Developing AI appropriately thus requires a multi- and inter-disciplinary approach.

Public trust in AI systems also plays a role in the spread of disinformation generated from AI generative software. This trust is rooted in how the public perceives the responsibility of AI for promoting socially good practices. Researchers that encourage the use of AI for these purposes contend that a system of ethics and public responsibility can be programmed into the software from the start (Dastani & Yazdanpanah, 2022). Therefore, the individuals behind the

[4] Although this study primarily examines the lack of regulations or outlined ethics surrounding AI in the EU, the reasoning behind it is relevant for this study, despite its focus on US-based companies.

development of AI must be at the center of the any research that hopes to point to the benefits of AI generative software.

Researchers against widespread AI use tend to gravitate toward its use in different sectors. Studies on emotional AI, for example, have recently begun to caution against implementing AI systems in public spheres. Emotional AI concerns the development of AI that attempts to mimic human emotion through body language, spoken language, and other nuances that communicate emotions to other people. Law enforcement and other public-sector organizations are interested in this development to assist in criminal investigations, but the evidence that this technology would assist is weak at best (Podoletz, 2022). Using AI to monitor emotional responses leads to "inferences and probabilistic predictions about…emotions and intentions" that could otherwise not be predicted (Podoletz, 2022). Researchers against the widespread implementation of AI argue that areas like emotional AI attempt to quantify human behavior, something that dangerously assumes intentions or emotions can be predicted with absolute certainty. Efforts to "rehumanize" algorithmic systems attempt to circumvent the problems associated with developing technology along such lines but operate in a manner that does not take in the perception of AI among the public (Ruckenstein, 2022).

More extreme takes on AI ask if it will end privacy or create a world that does not value human experiences. These concerns stem mostly from the fears of how unregulated AI can be used by organizations that do not have knowledgeable people at the helm (Walsh, 2022; Podoletz, 2022; Ruckenstein, 2022). Thus, AI development must consider the expertise and opinions of those that understand the ethical concerns behind it. Accordingly, research produced in industry (outside of academia) has begun to approach this question at a more rapid rate (Hagendoff & Meding, 2021). Recommendations based on principle for industrial considerations

– such as those associated with food safety – are becoming more popular, especially as it becomes clear that socio-technical approaches are part of the best practice for machine learning (Sapienza & Vedder, 2021). Before exploring the ways in which AI is integrating into society, it is necessary to examine the history and usage of the most popular programs currently on the market.

ChatGPT

OpenAI was founded in late 2015 by Elon Musk, Sam Altman and other Silicon Valley venture capitalists. In November 2022, the company released their software (ChatGPT) to a select few, then followed up with a public release in 2023 (Lock, 2022). ChatGPT is an LLM that allows users to "prompt" the AI to answer questions. Its accuracy in generating clear, largely correct content has made it a success in online spaces. It has become a problem in academia, where some students have used it to write short online responses or entire essays. While some information from ChatGPT is accurate, the current consensus from experts is ChatGPT can give accurate basic information but fails to complete complex tasks. When asked to generate a research paper, for instance, it will sometimes offer inaccurate information or give fake references in its bibliography or works cited.

ElevenLabs

ElevenLabs was founded in 2022 by Piotr Dabkowski and Mati Staniszewski. Dabkowski previously worked at Google as a machine learning engineer and Staniszewski worked at Palantir as a deployment strategist (*About ElevenLabs*, n.d.). ElevenLabs bills itself as "the most realistic AI speech software," allowing users to enter text prompts and hear it spoken aloud in humanlike voices. ElevenLabs also allows users to upload audio samples and "clone a voice" of an individual. The software has been used for some comedic purposes, but also in impersonation

attacks. It has also been used by security researchers to subvert voice identification and to spread misinformation.

MidJourney

MidJourney Incorporated was founded by David Holz in San Francisco, California. MidJourney released their software to open beta on July 12, 2022. It is used to generate AI created art. A user can input a prompt and it will generate a picture based on user input and learning from available images. MidJourney has caused some controversy among artists who fear it is a form of plagiarism. These concerns are not without merit – there are users creating AI images with the intent to sell them, potentially violating an artist's copyright.

Generative AI and other forms have created an artificial intelligence "arms race" between private companies, governments, and individuals. Every major technology company is now attempting to use existing artificial intelligence programs to cut back on employees or freeze hiring while also developing their own for internal purposes (Tangalakis-Lippert, n.d) .Worryingly, some companies (like Microsoft), while developing in-house AI programs, have cut their entire ethics team (Newton, 2023). As a result, instead of wonder and fascination, most AI news is filled with dread about lost jobs and livelihoods.

GitHub Copilot

GitHub Copilot was created by GitHub (owned by Microsoft) in collaboration with OpenAI and released to the public in October 2021. GitHub Copilot allows a programmer to use generative AI to assist them in writing code. A programmer can input a prompt in a natural language and the program will output code. GitHub Copilot's dataset was originally trained using OpenAI's, but Microsoft, much to the dismay of the community using Github, started using code from publicly and privately available code repositories for training. Licensing issues arose from

this action, with Microsoft, OpenAI, and GitHub ending up in legal disputes by parties affected by this “code stealing” and mixing (Claburn, n.d.).

Justification of Case Studies

This study will examine three cases in which generative AI seriously and negatively impacted a group of people. Case studies are not the most common means of research, namely because they point to a particular incident and extrapolate to larger assumptions. Due to the nature of social media, disinformation, and vastness of social engineering, case studies are needed in order to obtain a better understanding of events (Zeebaree et al., 2020). Additionally, due to the lack of access permitted to researchers by the companies running generative AI software, these case studies must be examined at a distance but through the lens of a security professional.

Since generative AI has little in the way of academic study, some of the reports of incidents across the United States may be factually questionable. This analysis will attempt to mitigate that problem by analyzing not only the event itself, but how the news is received by widely accessed websites. In examining the ways it is being reported, we can see how unmitigated access to AI is received by laymen. This analysis will also look at how the spread of disinformation and misinformation has exploded in the Information Age, which has arguably caused greater social isolation among people living in the US (Putnam, 2000). Therefore, the recommendations provided in later sections will center upon the need for education and restructuring of social expectations.

Below are brief overviews of the case studies that will be examined in other sections. Each summary comes from a major news reporting agency, with a brief explanation of why that story was chosen.

Case Study One: Explosion at the Pentagon

In late May 2023, an image was circulated on Twitter by a verified account showing a "confirmed" explosion at the Pentagon. Experts quickly identified the image as AI generated, but the damage was done. Per CNN's reporting, the stock market took a massive hit as speculators assumed the government was under some form of threat or in a moment of chaos. It quickly bounced back, but the prospect of an AI-generated image causing real-world damage stuck.

Little academic research has been done on the perception of the Internet and its impact on the "real world," particularly among adults in their 40s and 50s (CNN's primary demographic). The research that has been conducted examines how traditional university students and adolescents perceive the Internet in their studies, which has minimal bearing on this study. Among economists, there is speculation that the perception of the Internet as a separate entity from reality has created a level of "uncertainty" that could have significant implications for macroeconomics (Bontempi et al, 2019; Kurnia et al, 2006). Adopting a socio-technical approach to these case studies is essential, since public interaction is at the heart of analysis (Sartori & Bocca, 2023).

Case Study Two: Eating Disorders and a Chatbot

Shortly after the incident involving the Pentagon, the non-profit organization National Eating Disorders Association (NEDA) fired most of its staff in favor of hiring a chatbot. With goals of assisting as many people suffering from eating disorders as possible, the agency hoped to broaden their outreach efforts.[5] They adopted a chatbot named Tessa to recommend solutions to users in distress. Within 24 hours of its implementation, the tested bot began to offer harmful

[5] At the time, the staff in question were seeking to unionize to demand better working conditions and pay. There is some speculation that the use of AI was used as a scapegoat to prevent unionization of employees.

advice to its users. NEDA quickly pulled the chatbot to prevent further damage to people seeking aid.

The incident with Tessa mirrors a similar phenomenon that occurred with Tay, a chatbot released to Twitter in 2016 by Microsoft. Within hours of Tay's introduction, the bot began spewing hateful messages to the Twittersphere, leading to Microsoft pulling it after only 16 hours. Although amusing, researchers speculate that the public perception of the Internet and anthropomorphizing of the bot contributed to its downfall (Sartori & Bucca, 2023; Zemčík, 2021). The trust placed in Tessa to dispense accurate information in an ethical manner demonstrates a fundamental failing on public understanding of AI; these machines are first and foremost unable to think but are treated as though they can. They do not have a moral compass unless given one that aligns with its developer. Therefore, they will – without hesitation – produce and spread disinformation if it falls in line with the set goals.

Case Study Three: Professor and the AI Failed Class

At the end of the Spring 2023 semester, a professor at Texas A&M University-Commerce incorrectly accused his entire class of using ChatGPT for their final assignments. The professor – Dr. Jared Mumm – put each paper through ChatGPT and asked the software if it had written them. ChatGPT responded "yes" to Mumm's query, which led him to believe that each of the papers was written by the software. Mumm then erroneously assigned every student a zero on the assignment and let the students know they had all failed the course. He offered a makeup assignment, but the damage was done, and the story went national.

While Mumm's actions may seem extreme to most, he exemplifies the ignorance that surrounds AI among those not examining it closely. Additionally, it shows the failings that can occur when organizations do not have strong policies surrounding new technology. Trust in AI

and its abilities has surpassed that of individual trust in human-led organizations (Choung et al, 2022). This has led to an assumption that AI has inherent ethics, which – similar to the case involving NEDA – demonstrates the problems associated with not understanding AI as a machine.

The sources used for these case studies show how mass-use and availability of AI generative software has impacted the general public in different sectors. AI literature is still in its infancy, but it is clear that a multidisciplinary approach is needed in order to fully understand it. By adopting a socio-technical perspective, we can understand how social media access and social perception of AI is something that security professionals should be aware of (Zeebaree et al, 2020; Sartori & Bocca, 2023).

METHODOLOGIES

For this project, I primarily used keyword searches and gathered information from major news organizations regarding the effects of generative AI and disinformation. My initial search regarding news outlets utilized industry-based organizations and traditional news organizations (such as CNN and Rolling Stone). I chose the traditional news networks because they are the most likely to be seen by the general public. It is unlikely that someone unfamiliar with cybersecurity practices would seek out information within the industry. Thus, news regarding generative AI is mostly likely to reach the general public through mainstream news outlets. As interest grows in these subjects, more news is reported, and the relatively user-friendly interface used by generative AI platforms becomes more well-known, particularly to malicious actors. Therefore, although unconventional, the traditional news organizations' reporting of instances of disinformation provide a unique opportunity to examine how non-security professionals learn about new technologies that could affect the industry.

Finding literature for this project involved combing Google Scholar using keywords. These keywords included:

- generative AI
- generative artificial intelligence
- AI and disinformation
- generative AI and disinformation
- social engineering and AI
- mass disinformation and artificial intelligence
- legislation and artificial intelligence

These keywords provided a wealth of literature that covers the subject matter individually but does not combine the problems around disinformation and generative AI. The searches led to a reputable journal – *AI & Society* – which has been publishing studies on artificial intelligence and its possibilities since 1987. I then went through the journal, reviewing each issue for discussions on AI and its relationship with disinformation. Most of the research was multi- or inter-disciplinary, indicating that the case studies I had pulled from traditional news organizations should be viewed through multiple lenses.

One of the perspectives that the case studies needed to be viewed from comes from social engineering. A contentious topic among cybersecurity academics, social engineering is not something that can be clearly, objectively defined since it primarily deals with human behavior. The social sciences have long attempted to quantify human behavior, and security professionals have followed suit. Although difficult, based on the research conducted by sociologists and psychologists, we can make some accurate predictions on what a person will do in a given

scenario. The social engineering books used in this study are generic but wildly sold and purchased, demonstrating that they are perceived as valuable in the security realm. The practices described within have become ubiquitous within professional communities involved in cybersecurity, in turn showing malicious actors how an organization will behave. Therefore, it is vital to understand how cybersecurity as a discipline perceives social engineering.

DISCUSSION OF THE FINDINGS

Social engineering is generally viewed as an action taken against a specific person or organization for the purpose of gaining access to a network. These actions are ordinarily taken in the interest of profit for the person using social engineering. The goals of social engineering are thus not broad but narrow in scope. Disinformation does not fall into the study of this field because researchers do not interpret it as being related to the motivations behind social engineering. The ubiquity of Internet-connected devices, however, necessitates an expansion of the understanding of social engineering to include disinformation.

Disinformation targets a large group with the intent of spreading false information. These manufactured lies spread like a disease throughout online spaces, thriving in echo chambers of conspiracy. In some cases, false information actively causes harm to individuals.[6] Due to the scale of its damage and the methods used to employ disinformation, it has significant social impact. Therefore, disinformation should be included in studies of social engineering. Otherwise, cybersecurity professionals disregard a significant form of attack employed by malicious actors.

[6] For example, some disinformation about the COVID-19 vaccines was spread via social media that caused significant harm. Individuals skeptical of the vaccine viewed such disinformation as verification that they had a right to be nervous, and therefore chose not to receive a vaccination. This cause the development of herd immunity to slow, harming individuals that did not have the ability to get the vaccines.

Generative artificial intelligence is a tool, one which malicious actors have used to attempt social engineering and disinformation attacks. With the newest generative AI, malicious actors have been able to carry out various attacks that took time, effort, and skill. AI grants these actors the opportunity to conduct complex attacks more quickly.

Based on the research conducted on the case studies identified, unmitigated access to AI has resulted in a sharp increase in the spread of disinformation. Rapid development of Internet-based services without corresponding public education has contributed to the current problem. Most Internet users do not have the knowledge necessary to differentiate between AI generated content or something that is real. Additionally, mainstream media sources frequently accessed by the public have often sensationalized the rise of AI. These actions have not only created a sense of fear in the public regarding AI development but have also showed malicious actors the capabilities of it. Generative AI has made it easier for scammers and cybercriminals to take advantage of the fear and ignorance surrounding it, resulting in an increase in cases of harmful disinformation.

Case Study 1

The picture shown in appendix A shows an explosion at the Pentagon, which was circulated in late May 2023. Although the image was quickly determined to be generated by AI, it spread quickly across Twitter. The users that shared the image had a blue check mark, indicating they were “verified” by Twitter as legitimate users. Twitter verification is primarily viewed as a method to measure the reliability of a particular user. It also serves as a way to determine the social media accounts of celebrities or other important public figures, differentiating them from parody accounts. At the time of the incident, Twitter had changed its verification system to a paid model. This action hurt the legitimacy of the blue check marked

accounts, but not to casual users of the platform. Therefore, at a glance, it was difficult to determine whether information was coming from a reliable source.

As a result of users spreading the disinformation, it was acknowledged by news networks. The Indian news networks Republic TV picked up on the story and reported that the Pentagon had been attacked. Additionally, Russia Today (RT) mocked the agencies and users that took the attack seriously. Seemingly a ridiculous series of events, the incident had real-world consequences: the Dow Jones dipped 80 points and the S&P 500 dropped to down 0.15%. Both stocks recovered within minutes of the truth coming out, but it remains a cautionary example of the real-world implications of disinformation.

Case Study 2

Disinformation can come in the form of not only blatantly incorrect information, but actively harmful to individuals. In May 2023, the non-profit organization NEDA turned to a chatbot (Tessa) as replacement for its staff. They hoped that a chatbot could provide adequate advice to their clientele (the majority of whom suffered from eating disorders). A cost-saving measure, the chatbot quickly proved itself incapable of providing adequate advice, offering harmful recommendations to NEDA's patients. In this case, the disinformation was spread by the AI itself, demonstrating that oversight is needed with AI implementation.

Poor decision making aside, the actions of NEDA demonstrate an inherent trust of AI. This trust is built from the false assumption that an AI comes with a built-in sense of morality, similar to most people. Research conducted by Joseph Weizenbaum in the 1960s regarding chatbots demonstrated that a person unfamiliar with the technical aspects of machine learning will treat that machine as though they are a person. Dangerously, the way organizations interpret potential uses of contemporary AI shows that they are unwilling to acknowledge that the AI does

not come with a system of ethics. It is the task of the developer to input ethics into a machine, otherwise it will adopt what is around it, invariably mimicking the worst of humanity when unleashed upon the Internet (Sartori & Bucca, 2023; Zemčík, 2021).

Incorporating ethics into machine learning is a point of contention among professionals working with the technology. The issue becomes more contentious as computer security professionals' concerns regarding its use in potential attacks becomes relevant. Tessa exemplifies how a chatbot can spread dangerous disinformation with the legitimacy of an established organization backing it.

Case study 3

In academia, generative AI has become a controversial subject. Universities are scrambling to establish general guidelines for instructors and professors to follow, especially given the rampant use of AI to skirt assignments. In the Spring 2023 semester, Dr. Jared Mumm of Texas A&M University-Commerce demonstrated how ignorance and sensationalized stories can lead to damage in real-world instances. For a final assignment, Mumm put each of his student's papers through ChatGPT, asking it if it had written it. True to form, ChatGPT answered "yes" to each paper, despite not having written them. Mumm accused his entire class of using AI to write their paper, assigning zeros to everyone in the class.

Similar to the incident with Tessa, this incident exemplifies the problems regarding trust in the information AI presents. In this case – rather than assume that the AI had a human sense of right and wrong – Mumm assumed that the AI would not erroneously claim to have written the essays. He placed a trust in the machine's ability to always be correct, neglecting that there is a human aspect to the creation of machine learning. As of writing, Mumm's actions demonstrate that trust in AI's abilities has surpassed that of humans. Therefore, should an AI generate

disinformation, it could spread more easily through the public since they view it as inherently trustworthy.

The incidents described demonstrate a problem with how the public views AI and its ability to be "wrong." There is a prevalent belief that a machine cannot be wrong, in that it cannot produce an answer that is incorrect based on the information given. If a machine generates an incorrect answer, it is assumed that the human user behind the initial input made an error. This assumption extends to ideas on what is right and wrong in a moral sense. Users assume that the machine has an inherent understanding of what is right and wrong, which leads them to trust its decision making in various circumstances.

Generative AI has made the spreading of disinformation trivial. For doctored photos, MidJourney, some clever prompts, and a social media platform to push the images created on is all that is needed. Likewise, for faked audio, a subscription to ElevenLabs and audio clips from the targeted voice – allow easy cloning of a voice. These services are frequently used to create disinformation. In MidJourney's case, an unpracticed artist can easily input a prompt and get a picture of whatever they want.

Generative AI is widely available to the public. Most of the generative AI tools mentioned in this paper previously were free, but now cost money, restricting some usage. Below are the details of their cost and access. These breakdowns demonstrate the level of access available, thereby showing the threat they present to organizations and society at large.

ChatGPT

Costs are based on datasets used, and type of output wanted. ChatGPT uses a token system, in which tokens are available for purchase in exchange for additional services. For example, 1,000 tokens is about equivalent to 750 words. A small amount of free tokens are given out for

"research purposes," but the prompt is limited to GPT-3.5. Services offered by GPT-4 are limited to paying individuals.

MidJourney

MidJourney has several plans including a free tier, with each providing more resources. Higher plans allow private art remixing – art not shared with the public. Plans range from free trial, $10, $30, $60 per month, respectively.

ElevenLabs

ElevenLabs currently has six different plans, including a free trial. Plans range from $5 - $330 a month, which allow varying levels of access. The free trial does not allow commercial monetization and requires attributing anything made to ElevenLabs. The cheapest monthly plan ($5) gives access to a feature called "voice cloning." Voice cloning allows a user to take up to twenty-five samples of a given voice, then make a "clone" of it. With that clone, the generated voice can be made to say virtually anything. Additionally, under this plan, users can use this cloned voice for monetization and do not have to attribute it to ElevenLabs. Other than an ethics and terms of service, ElevenLabs does not appear to stop users from uploading samples of users without their consent, allowing malicious actors to attempt numerous social engineering and disinformation attacks.

For malicious actors, they will act maliciously when it comes to spreading disinformation and attempting social engineering attacks. For non-malicious end users, some training should be given on how to use generative artificial intelligence ethically and legally. For the companies that develop generative AI, considering most of them laid off their ethics teams, I argue they are the most responsible for this current wave of disinformation, misinformation, and a steady increase

in social engineering attacks. They put profits above all else – which means any ethical constraints go out the window in the ever-fragmented generative AI landscape.

ElevenLabs knows that their software is being used to spread disinformation, and instead of slowing down the number of accounts they accept, or restricting voice cloning tools, they provided a one-page document on acting ethically. As far as this author knows, there is no technical or administrative controls that ElevenLabs applies to their software to stop these types of voice cloning attacks. ElevenLabs primarily puts it on the end user to act ethically. While it may give some users pause and reevaluate what they are doing, malicious actors may not have that level of care.

RECOMMENDATIONS

The recommendations from this study can be separated into four categories. First, education for end users of generative AI – including in university or other academic and professional settings – must become a priority for AI developers. Second, legislation from state institutions that advocates for protection of intellectual property and the privacy of its people must be prioritized. Third, further protections for creative individuals can be provided through the establishment and effective management of artistic trade groups, professional organizations, unions, and collective bargaining units (CBUs). Lastly, the developers of AI must establish clearer ethical guidelines for their products. These ethical guidelines can then be adapted by professional and academic organizations to help meet the demand AI has created. Although these recommendations may adopt a fatalistic attitude toward AI, I believe it is imperative that such actions occur to ensure the safety and privacy of citizens in the United States.

Education

End users of generative artificial intelligence should be educated on the limitations of generative AI, specifically concerning ChatGPT. Understanding the mechanisms behind ChatGPT and other chatbots will help users to understand what they are getting out of the service. This can help mitigate the belief that everything that is generated from the software is infallible, or that it follows particular ethical principles. These actions seek to undo generational beliefs in machine perfection. In doing so, it could improve the perception of the relationship between ML software and end users.

Local universities, community colleges, community centers, and libraries can play a part in educating on generative AI. Numerous studies have shown that a well-educated populace is more resistant to disinformation, misinformation, and social engineering attacks (Hwang et al., 2021; Adjin-Tettey, 2021). Growing resentment toward education institutions, including a mistrust for academics and science-based fields, has led to public disregard for the importance of education (Brenan, 2023). Creating an environment that is friendly toward education would help mitigate this problem. A friendly environment toward education means providing more funding and offering more accessible methods of education for the general population.

Legislation of Artificial Intelligence

The United States must pass legislation addressing the unmitigated use of AI. In AI's current form, large corporations and individual users are able to use it to hoover up vast amounts of copyrighted work, remix it, or use it to train datasets. Although some arguments exist that these actions are not violating the copyrighted works or intellectual property of the artist, it has caused significant anxiety within creative communities regarding their value.[7] Rampant,

[7] In the summer of 2023, the Writers' Guild and Screen Actors' Guild – American Federation of Television and Radio Artists initiated strikes in response to studios underpaying for their talents. Major

unregulated use of AI is outlining a clear path to job insecurity and mass disinformation events. By offering regulations (informed by experts in generative AI), state governments can prevent potential problems that could arise. Additionally, state legislation could promote education in the area, allowing for greater multidisciplinary research to be conducted on the development of AI.

Trade Groups, Professional Organizations, Unions, Collective Bargaining Units

Trade groups, professional organizations, unions, and CBUs that relate to creative disciplines or other fields related to AI need to establish more effective and clear guidelines for their members. Developing contracts that protect copyrighted work and intellectual property will provide safeguards for creative professionals. In the event that an artist's work is used to spread disinformation or misinformation, such protections would prevent the original artist from facing backlash. Additionally, it would protect their work from plagiarism.

These groups can also assist in ensuring that artists are compensated appropriately for their work that is used to train AI. Ignorance has driven creative conglomerates to attempt to enlist the aid of AI as a cost-saving measure in creative productions. Unions and CBUs would assist in protecting the individuals in the industry from being pushed out in favor of machine-generated work.

Ethics in Computer Science and Cybersecurity

For aspiring computer scientists, universities must increase efforts at integrating ethics and humanities classes for students within their curriculum. Single ethics courses often do not have enough time to cover the breadth of incorporating morals into future computer science work. By placing more emphasis on these elements of education, computer scientists can understand the importance of their work and the implications of operating without ethics.

---

studios then began to release statements related to using AI to replace writers and other creatives, intensifying the strike.

Therefore, ethics courses should extend to curriculum in other areas of the classroom, placing greater attention on its importance.

If an aspiring computer scientist finishes their degree and develops technology that potentially alters humanity without appropriate care, then the university and educational systems at play have failed. Creating new technologies and releasing them to the public without ethical considerations or the input of persons familiar with the potential socio-economic impact demonstrates a fundamental misunderstanding of computer science ethics. Improving education in ethics and requiring coursework in related fields would help broaden aspiring computer scientists' and security professionals' perspectives.

Incorporating greater ethical considerations creates room for more safeguards. Developers would have the tools necessary to make new technologies – such as AI – less accessible to malicious actors. Additionally, ethical considerations would motivate computer scientists to assist in end user education, thus mitigating the spread of disinformation. The recommendations provided focus on reconsidering the development of AI and the way it is introduced to the public. Evidence suggests that creating a new technology and unleashing it without proper consideration creates problems in nearly every sector of society. Security professionals – working alongside computer scientists – have the opportunity to create better technologies that protect the lives and privacy of individuals and organizations.

CONCLUSION

At the time of writing, generative AI has become a point of contention in nearly every sector of society. Among businesses, it is being used as both a tool and a point of condemnation. Its potential use in entertainment has contributed to a rise in worker resistance to unfair working conditions and practices. Among educators, it remains a consistent issue as students use it to skirt

their learning responsibilities. As it enters these different sectors and becomes something that the general public is aware of, it increases in popularity and use. Most generative AI functions by "learning" from user input. Thus, the more often it is reported on and used, the more robust it becomes.

Although AI has a multitude of positive uses, its increasing popularity has made it a tool for malicious actors. Those that wish harm upon others utilize the services provided by AI to access secure networks or demand a ransom for held valuable information. In viewing these instances as reported by significant media outlets, these instances beg the question: How has AI generative software impacted the frequency and ferocity of disinformation attacks? This research has examined the ways in which unmitigated access to and frequent aggrandizing of AI has led to it turning into a tool for malicious actors.

This study presented the hypothesis that increased reporting on and access to AI had created an environment that saw increased social engineering attacks and mass disinformation events. With the ability to easily mask identity and maintain anonymity, the risk of carrying out such attacks is considerably lessened. Therefore, malicious actors have adapted to AI and begun to use it for their own nefarious purposes. The case study research in this thesis has presented evidence that – while AI can be used for good – it has frequently resulted in mass disinformation and social engineering attacks that have real world consequences. Due to its availability, mass disinformation attacks are now much easier to conduct, which has therefore increased their frequency. Additionally, due to the "evidence" that can be created to accompany instances of disinformation, it makes it more believable to the users that see it at a glance.

Current literature covering the effects of AI on disinformation and social engineering is virtually nonexistent. Due to this lack in research, the literature for this study focused on three

primary domains: social engineering, disinformation, and generative AI. Social engineering is a subject that has generated significant controversy and spans multiple disciplines. Within network security, the study of it mostly focuses on ways to mitigate its impact on profit-oriented organizations. Social engineering is also analyzed to construct training for people unfamiliar with the discipline so they may learn to prevent malicious actors from gaining access to critical data. Outside of network security, social engineering is analyzed primarily by its motivations. Sociologists and psychologists examine the ways people are manipulated by it, and what they can learn about human experiences by studying its effects. Due to the varied perspectives on social engineering, it is important to take a multidisciplinary approach when conducting research on it.

Disinformation has a significant amount of prior research done on it, especially in the wake of the 2020 US presidential election and global events surrounding COVID-19. As such, research on its effects has also spanned across disciplines. Security professionals have tended to examine disinformation from a standpoint that reflects their approach to social engineering. Disinformation can cause significant harm to an organization or individual, and usually requires insider information to be spread effectively. Therefore, using techniques similar to those used to mitigate social engineering, professionals can do the same with disinformation attacks.

Disinformation and social engineering are closely related but studied together infrequently. Security professionals generally do not view disinformation as warranting the same amount of attention. As malicious actors use new technologies to increase the effectiveness of their disinformation attacks, it becomes more apparent that security professionals need to take it seriously. Therefore, based on the literature examined in these two areas, I conclude that

disinformation and social engineering should be addressed in a multidisciplinary and similar way to protect private and public assets.

Generative AI fell into the public scene in the later months of 2022. Thus, the literature surrounding it is relatively sparse. The academic journal *AI & Society* – founded in 1987 – had previously examined the prospect of AI use in modern society. Generative AI has brought new research to academic areas as researchers scramble to understand its implications or potential uses. Most of the publications in *AI & Society* examine the ways AI can be used in niche circumstances. For example, Chu et al (2020) explore the ways that AI could be used to better analyze radiology screenings. The flipside of this research has a more concerned tone regarding AI. These aspects of research – from academics like Mehdi Dastani and Vahid Yazdanpanah (2023) – identify the problems that unmitigated access to AI can bring. Their warnings fall mostly upon deaf ears, as increasing reports of AI usage for unethical practices increases.

To effectively interpret the effects of generative AI, this research looked at three case studies. Although these case studies do not encapsulate the entirety of a person's experience with AI, they demonstrate its capabilities, effect in real-world areas, and the reaction to them from mainstream news networks. The recent development of AI makes it difficult to conduct a complete analysis on its uses. Additionally, the organizations that have released the software are reluctant or unwilling to allow researchers to see collected data. Therefore, academics are reduced to using what is available – primarily reported-on instances of disinformation using AI and the reaction by the general public. These case studies have provided the backdrop of this research, demonstrating that dangerous use of AI is not only frequently reported on, but is becoming a greater presence in different sectors of society.

The first case study, as outlined in previous sections, concerned the faked photograph of an explosion at the Pentagon. The photo generated significant media buzz, causing stock prices to fall in the New York Stock Exchange. This form of disinformation was generated using MidJourney, an AI capable of generating pictures based on select prompts. Although it had a real-world, negative impact, it is important to consider that the image may not have been created with the intention to spread disinformation. Rather, it may have begun as a joke, but spread quickly through social media, uninhibited. Therefore, it has become easier for users to exploit AI in ways that have harmful consequences.

On a smaller scale, the second case study addressed the incorrect information fed to eating disorder patients from a chatbot employed by the nonprofit NEDA. This study demonstrated an issue that runs rampant through AI usage: inherent public trust. The general public trusts machines to know what to do without error, which extends to understanding when something is considered unethical. Blind trust in AI has resulted in people considering them infallible, failing to recognize that they are the products of man. Generative AI only has the ability to be as ethical as it is programmed to be. Without oversight, it can and will adopt the ethics of those it learns from, as evidenced by Tessa and Microsoft's Tay (Sartori & Bucca, 2023; Zemčík, 2021).

Inherent trust in AI extends into the third and final case study, where Dr. Jared Mumm erroneously failed his entire class, believing their papers were written by ChatGPT. He ran the papers through ChatGPT, asking the software if it had written them. The AI claimed it had, which Mumm believed, leading to him failing his class. The prospect of the AI being incorrect did not occur to Mumm during the process of checking his student's papers. Thus, his actions

demonstrate the way that disinformation can easily spread through the use of AI; trust in the system to perform appropriate and truthful actions leads to trust in all information output.

Creating instances of disinformation had previously been a relatively labor-intensive task. With unmitigated access to AI, malicious actors can now create images, text, or voice recreations of nearly anything. Since these are produced with convincing accuracy, the information can be spread quickly without much resistance. Additionally, public trust in machine ability to produce accurate, cohesive information leads to little questioning of the material produced by AI. Therefore, the disinformation problem produced by generative AI is twofold. First, malicious actors may use it to create false information for the sake of humor or personal gain, taking advantage of what little attention online users generally pay to the material in their social media feeds. Secondly, due to inherent public trust in information produced by machines, disinformation produced by AI is believed outright. This leads to individuals mistakenly believing anything that is churned out, making the AI itself a harbinger of disinformation.

Based on the information learned in the case studies, the initial hypothesis for this research was only partially correct. Increased reporting on and access to generative AI has led to greater use of the software, although not directly as social engineering attacks and mass disinformation events. Rather than the disinformation being intentionally created and distributed by malicious actors, the AI itself is producing false results and morally questionable material. Public trust in machines leads individuals to believe whatever is produced from them, which may or may not be false. Additionally, the ethics and morals that an AI possesses result from the creator – not the machine itself. Its moral campus is only as great as the creator saw to make it. Without public education or knowledge of this fact, information produced by AI will continue to become a problem, resulting in more mass disinformation events.

Increased reporting on generative AI has thus created a paradox. The more often instances of disinformation are reported on, the more likely a person with malicious intent is to utilize it to carry out their aims. If the instances are ignored, then AI could still be used to produce disinformation, whether intentionally or not. Therefore, one of the most viable solutions for slowing the impact of disinformation is education for end users of generative AI. Individuals interested in using the software must become aware of its potential impacts and how to use it ethically. Additionally, the creators of the software need to implement stronger ethical parameters into its design to prevent malicious users. As most security professionals are aware, though, no amount of protection can prevent the ingenuity of someone determined to cause as much damage as possible. Therefore, the best way to protect the public is to utilize the educational tools already in place.

A way that AI use can be mitigated (and protect industries) is through strict legislation. If a state entity is willing, they may initiate restrictions on the production of AI, thereby reducing the potential for public harm. Unfortunately, this would require significant resources invested in the study of the impact of AI and a multidisciplinary approach. Due to the political climate at the time of writing, this level of cooperation and dedication would be unlikely, barring a significant attack with AI at the head. International bodies could perhaps address the issue.

APPENDICES

Appendix A: Generative AI Photos

Figure a Dolan, Leah. (March 29, 2023). Look of the Week: What Pope Francis' AI puffer coat says about the future of fashion. CNN. https://www.cnn.com/style/article/pope-francis-puffer-coat-ai-fashion-lotw/index.html.

Figure b UKR Report [@UKR_Report]. (May 22, 2023). *#BREAKING An explosion was reported near the Pentagon* [Image Attached] [Tweet]. Twitter. https://twitter.com/N_Waters89/status/1660651721075351556/photo/1.

Appendix B: Gradient of AI Distribution

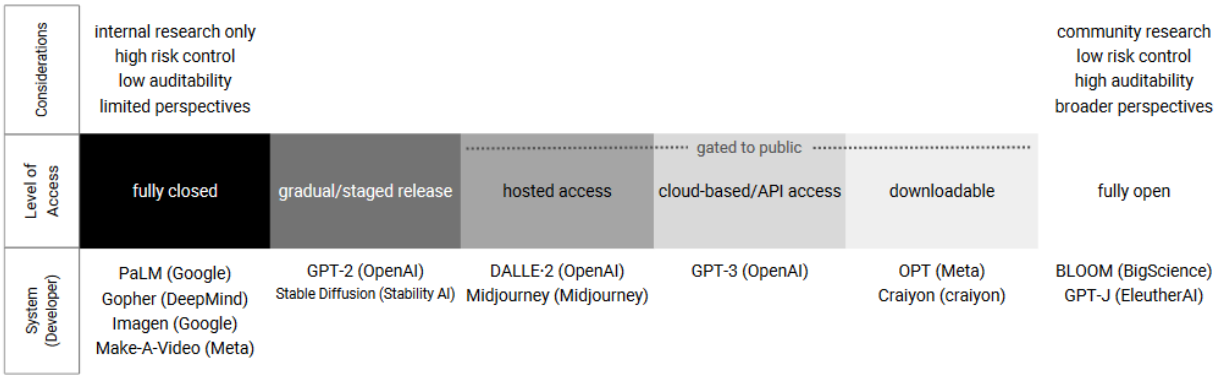

Figure 1, retrieved from Solaiman, I. (2023). The Gradient of Generative AI Release: Methods and Considerations (arXiv:2302.04844). arXiv. https://doi.org/10.48550/arXiv.2302.04844

Appendix C: Charts

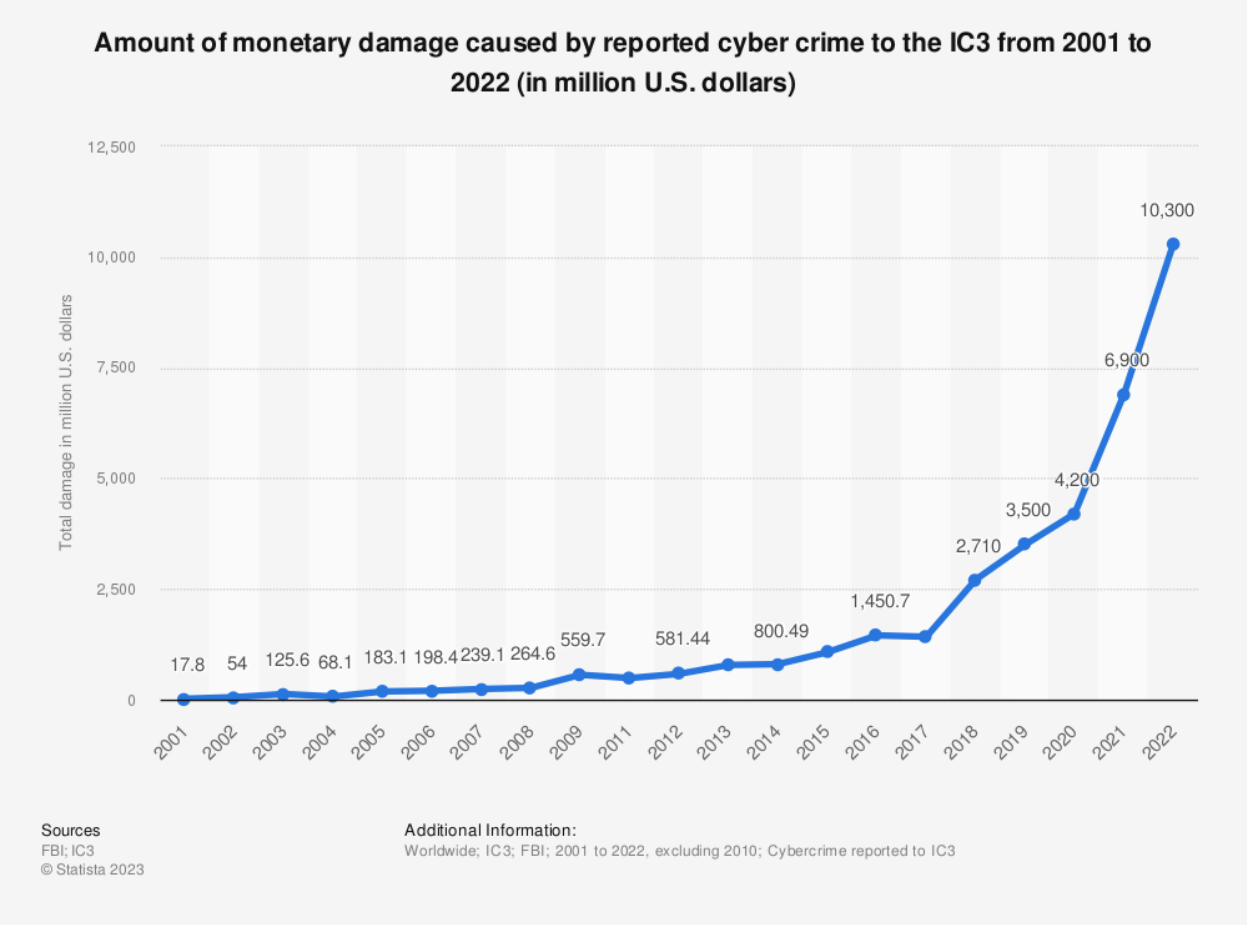

Figure c Source: Petrosyan, A. (2023, April 24). IC3: total damage caused by reported cyber crime 2001-2022. *https://www.statista.com/statistics/267132/total-damage-caused-by-by-cyber-crime-in-the-us/.*

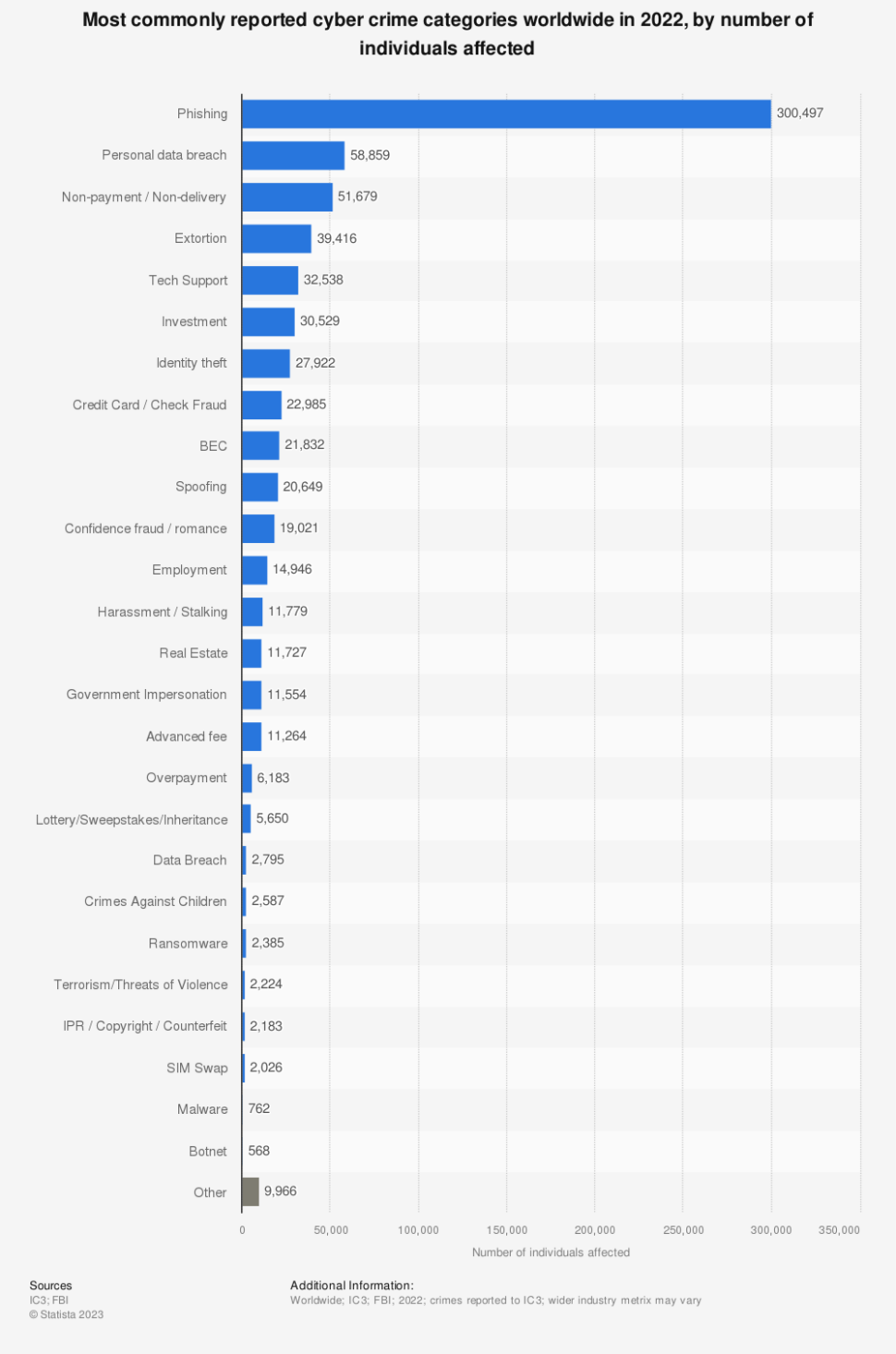

Figure d Source: Petrosyan, A. (2023, April 24). Most repeated types of cyber crime worldwide 2022, by number of individuals affected. *https://www.statista.com/statistics/184083/commonly-reported-types-of-cyber-crime-global/*

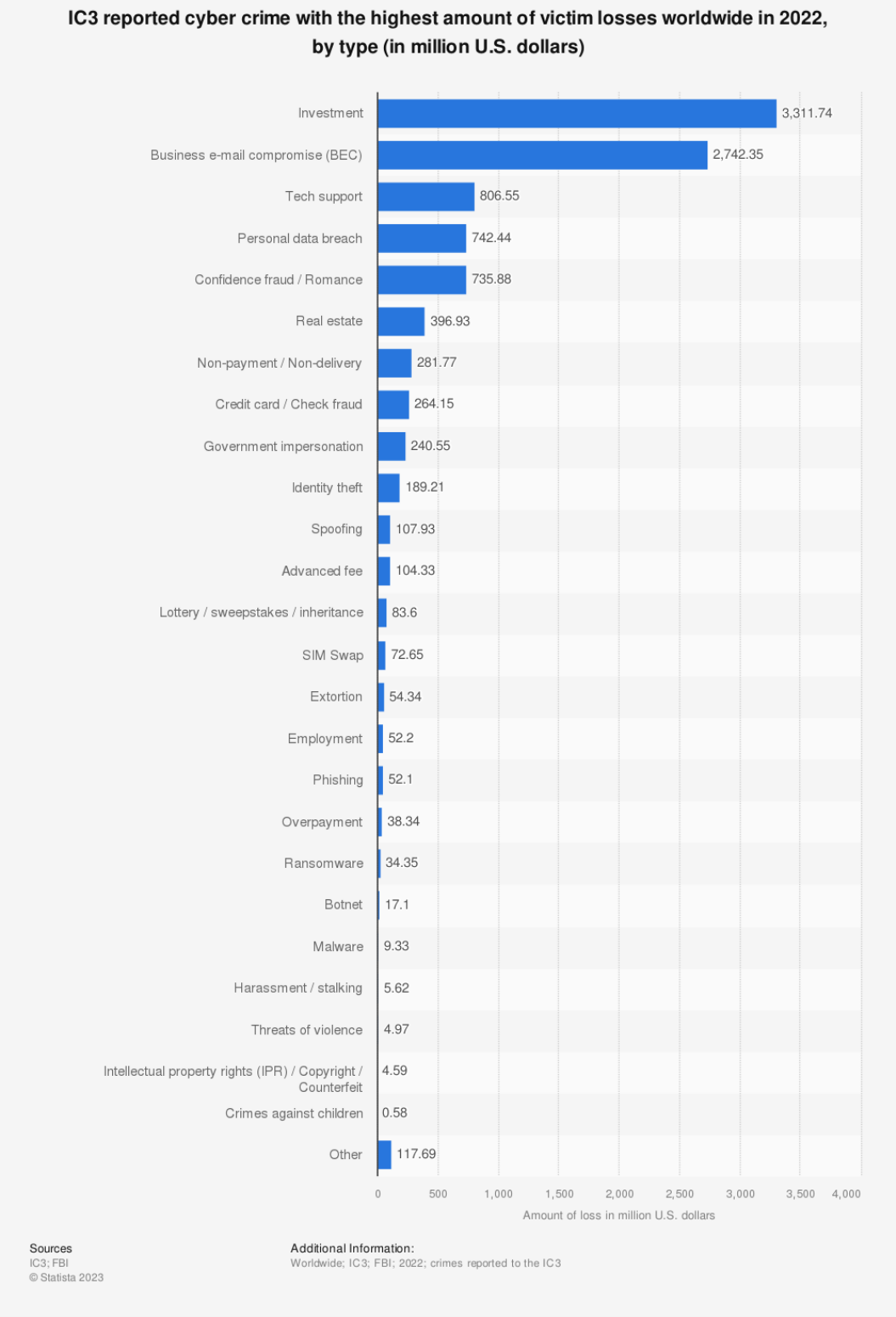

Figure e Source: Petrosyan, A. (2023, April 24). Leading cyber crime victim loss categories worldwide 2022. *https://www.statista.com/statistics/234987/victim-loss-cyber-crime-type/*

References

About ElevenLabs. (n.d.). Retrieved May 12, 2023, from https://beta.elevenlabs.io/about

Andrews, C. (2021-09). Fake news is not 'news'—It's manipulative disinformation. *Engineering & Technology, 16*(8), 1–8. https://doi.org/10.1049/et.2021.0807

Aroyo, A. M., Rea, F., Sandini, G., & Sciutti, A. (2018). Trust and Social Engineering in Human Robot Interaction: Will a Robot Make You Disclose Sensitive Information, Conform to Its Recommendations or Gamble? *IEEE Robotics and Automation Letters, 3*(4), 3701–3708. https://doi.org/10.1109/LRA.2018.2856272

Bontempi, M., Frigeri, M., Golinelli, R., & Squadrani, M. (2019). Uncertainty, Perception and the Internet. *SSRN Scholarly Paper*. 1-95. https://doi.org/ 10.2139/ssrn.3469503

Borch, C. & Hee Min, B. (2022). Toward a sociology of machine learning explainability: Human-machine interaction in deep neural network-based automated trading. *Big Data & Society 9*(2), 1-13. https://doi.org/ 10.1177/20539517221111361.

Chatbots: A Brief History Part I - 1960s to 1990s. (n.d.). Retrieved May 12, 2023, from https://www.botsplash.com/post/chatbots-a-brief-history

Choung, H., Prabu, D., & Ross, A. (2023). Trust and ethics in AI. *AI & Society 38*(2), 733-745. https://doi.org/ 10.1007/s00146-022-01473-4.

Chu, L., Anandkumar, A., Shin, H., & Fishman, E. (2020, April 17). The Potential Dangers of Artificial Intelligence for Radiology and Radiologists. *Journal of American College of Radiology 17*(10), 1309-1311. https://doi.org/10.1016/j.jacr.2020.04.010.

Claburn, T. (n.d.). GitHub and OpenAI fail to wriggle out of Copilot lawsuit. Retrieved May 13, 2023, from https://www.theregister.com/2023/05/12/github_microsoft_openai_copilot/

Cox, J. (2023, February 23). How I Broke Into a Bank Account With an AI-Generated Voice. Vice. https://www.vice.com/en/article/dy7axa/how-i-broke-into-a-bank-account-with-an-ai-generated-voice

Dastani M. & Yazdanpanah, V. (2023). Responsibility of AI Systems. *AI & Society 38*(2), 843-852. https://doi.org/ 10.1007/s00146-022-01481-4.

Diaz Ruiz, C., & Nilsson, T. (2023-01). Disinformation and Echo Chambers: How Disinformation Circulates on Social Media Through Identity-Driven Controversies. *Journal of Public Policy & Marketing, 42*(1), 18–35. https://doi.org/10.1177/07439156221103852

Fruhlinger, J. (2023, March 10). What is generative AI? The evolution of artificial intelligence. InfoWorld. https://www.infoworld.com/article/3689973/what-is-generative-ai-the-evolution-of-artificial-intelligence.html

Gehl, R. W. (2014). Teaching to the Turing Test with Cleverbot. *Transformations: The Journal of Inclusive Scholarship and Pedagogy 24*(1-2), 56-66. https://doi.org/10.5325/trajincschped.24.1-2.0056.

Göranzon, B., Florin, M., & Sällström, P. (1988). The concept of dialogue. *AI & Society, 2*, 279-286. https://doi.org/10.1007/BF01891362

Gray, J. (2022). *Practical Social Engineering: A primer for the ethical hacker*. No Starch Press.

Grinbaum, A., & Adomaitis, L. (2023). Dual Use Concerns of Generative AI and Large Language Models (arXiv:2305.07882). arXiv. https://doi.org/10.48550/arXiv.2305.07882

Guo, B., Ding, Y., Sun, Y., Ma, S., Li, K., & Yu, Z. (2020-11-11). The mass, fake news, and cognition security. *Frontiers of Computer Science, 15*(3), 1-13. https://doi.org/10.1007/s11704-020-9256-0

Guo, Z., Valinejad, J., & Cho, J.-H. (2022-09). Effect of Disinformation Propagation on Opinion Dynamics: A Game Theoretic Approach. *IEEE Transactions on Network Science and Engineering, 9*(5), 3775–3790. https://doi.org/10.1109/TNSE.2022.3181130

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other Large Generative AI Models (arXiv:2302.02337). arXiv. https://doi.org/10.48550/arXiv.2302.02337

Hadnagy, C. (2018). *Social Engineering: The science of human hacking*. Wiley Publishing.

Hagendoff, T. & Medding, K. (2023). Ethical considerations and statistical analysis of industry involvement in machine learning research. *AI & Society 38*(1), 35-45. https://doi.org/10.1007/s00146-021-01284-z.

Huang, K. (2023, April 8). Why Pope Francis Is the Star of A.I.-Generated Photos. The New York Times. https://www.nytimes.com/2023/04/08/technology/ai-photos-pope-francis.html

Johnson, J. (1988) Mixing Human and Nonhumans Together: The Sociology of a Door-Closer. *AI & Society 35*(3), 298-310. https://doi.org/10.2307/800624.

Kumar, D. (2023, March 30). Is ChatGPT causing layoffs? Which jobs are in danger due to ChatGPT? All FAQs. Mint. https://www.livemint.com/technology/tech-news/is-chatgpt-causing-layoffs-which-jobs-are-in-danger-due-to-chatgpt-all-faqs-answered-11680177842752.html

Lock, S. (2022, December 5). What is AI chatbot phenomenon ChatGPT and could it replace humans? The Guardian. https://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans

Mitnick, K & Simon, W. (2002). *The Art of Deception: Controlling the human element of security*. Wiley Publishing.

Myles, D., Benoit-Barne, C., & Millerand, F. 'Not your personal army!' Investigating the organizing property of retributive vigilantism in a Reddit collective of websleuths. *Information, Communication & Society 23*(3), 317-336. https://doi.org/10.1080/1369118X.2018.1502336.

Newton, C. (2023, March 14). Microsoft lays off team that taught employees how to make AI tools responsibly. The Verge. https://www.theverge.com/2023/3/13/23638823/microsoft-ethics-society-team-responsible-ai-layoffs

Nuki, T. (1990). The 'transfer of skill' and the 'transfer of human relations' to machine systems. *AI & Society, 4,* 173-182. https://doi.org/10.1007/BF01889938.

Prasad, P. (1995). Working with the "smart" machine: computerization and the discourse of anthropomorphism in organizations. *Studies in Cultures, Organizations and Societies 1*(2), 253-265. https://doi.org/10.1080/10245289508523458.

Podoletz, L. (2023). We have to talk about emotional AI and crime. *AI & Society 38*(3), 1067-1082. https://doi.org/10.1007/s00146-022-01435-w.

Putnam, R. (2000). *Bowling Alone: The collapse and revival of American community*. Simon & Schuster.

Ruckenstein, M. (2023). Time to re-humanize algorithmic systems. *AI & Society 38*(3), 1241-1242. https://doi.org/10.1007/s00146-022-01444-9

Salkowitz, R. (n.d.). Midjourney Founder David Holz On The Impact Of AI On Art, Imagination And The Creative Economy. Forbes. Retrieved May 12, 2023, from https://www.forbes.com/sites/robsalkowitz/2022/09/16/midjourney-founder-david-holz-on-the-impact-of-ai-on-art-imagination-and-the-creative-economy/

Sapienza, S. & Vedder, A. (2023) Principle-based recommendations for big data and machine learning in food safety: the P-SAFETY model. *AI & Society 38*(1), 5-20. https://doi.org/10.1007/s00146-021-01282-1.

Sartori, L., & Bocca., G. (2023). Minding the gap(s): public perceptions of AI and socio-technical imaginaries. *AI & Society 38*(2), 443-458. https://doi.org/10.1007/s00146-022-01422-1.

Sharma, K., Ferrara E., & Liu Y. (2022). Characterizing Online Engagement with Disinformation and Conspiracies in the 2020 U.S. President Election. *Proceedings of the Internetional AAAI Conference on Web and Social Media 16*, 908-919. https://doi.org/10.1609/icwsm.v16i1.19345

Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T. H., Ding, K., Karami, M., & Liu, H. (2020-08-15). Combating disinformation in a social media age. *WIREs Data Mining and Knowledge Discovery 10*(6), 1-39. https://doi.org/10.1002/widm.1385

Sjouwerman, S. (n.d.). Social Engineering Attacks Utilizing Generative AI Increase by 135%. Retrieved May 13, 2023, from https://blog.knowbe4.com/generative-ai-social-engineering-attacks

Smithers, T. (1988) Product creation: An appropriate coupling of human and artificial intelligence. *AI & Society, 2*, 341-353. https://doi.org/10.1007/BF01891367.

Solaiman, I. (2023). The Gradient of Generative AI Release: Methods and Considerations (arXiv:2302.04844). arXiv. https://doi.org/10.48550/arXiv.2302.04844

Sunstein, C.R., & Vermeule, A. (2009) Conspiracy theories: Causes and cures. *Journal of Political Philosophy, 17*(2), 202-207. Wiley Online Library.

Tangalakis-Lippert, K. (n.d.). IBM halts hiring for 7,800 jobs that could be replaced by AI, Bloomberg reports. Business Insider. Retrieved May 13, 2023, from https://www.businessinsider.com/ibm-halts-hiring-for-7800-jobs-that-could-be-replaced-by-ai-report-2023-5

Turchin, A. (2019, March). Assessing the future plausibility of catastrophically dangerous AI. *Futures 107*, 45-58. https://doi.org/10.1016/j.futures.2018.11.007.

What is Generative AI? Everything You Need to Know. (n.d.). Enterprise AI. Retrieved May 11, 2023, from https://www.techtarget.com/searchenterpriseai/definition/generative-AI

What Is Social Engineering—The Human Element in the Technology Scam| Cybersecurity | CompTIA. (n.d.). Default. Retrieved May 11, 2023, from https://www.comptia.org/content/articles/what-is-social-engineering

Who Are Hackers—The Testimony Of An Ex-Hacker | Hackers | FRONTLINE | PBS. (n.d.). Retrieved May 29, 2023, from https://www.pbs.org/wgbh/pages/frontline/shows/hackers/whoare/testimony.html

Walsh, T. (2022) Will AI end privacy? How do we avoid an Orwellian future. *AI & Society 38*(3), 1239-1240. https://doi.org/10.1007/s00146-022-01433-y

Yudkowsky, K. (2008). Artificial Intelligence as a positive and negative factor in global risk. In N. Bostrom & M. M. Ćirković (Eds.), *Global Catastrophic Risks* (pp. 308-346). Oxford University Press.

Zeebaree, S., Ameen., S. & Sadeeq, M. (2020). Social Media Networks Security Threats, Risks and Recommendations: A Case Study in the Kurdistan Region. *International Journal of Innovation 13*(7). 349-365.

Zemčík, T. (2021) Failure of chatbot Tay was evil, ugliness and uselessness in its nature or do we judge it through cognitive shortcuts and biases? *AI & Society 36*(1), 361-367. https://doi.org/ 10.1007/s00146-020-01053-4.

Züger, T & Asghari, H. (2023) AI for the public. How public interest theory shifts the discourse on AI. *AI & Society 38*(2), 815-828. https://doi.org/10.1007/s00146-022-01480-5.

Case Study References

Cohen, R. (2023, May 22). *An apparently AI-generated hoax of an explosion at the Pentagon went viral online – and markets briefly dipped.* Insider. https://www.insider.com/ai-generated-hoax-explosion-pentagon-viral-markets-dipped-2023-5

Klee, M. (2023, May 17). *Professor Flunks All His Students After ChatGPT Falsely claims It Wrote Their Papers.* RollingStone. https://www.rollingstone.com/culture/culture-features/texas-am-chatgpt-ai-professor-flunks-students-false-claims-1234736601/.

Morris, C. (2023, May 31). *National Eating Disorder Association shuts down A.I. chatbot it planned to use to replaces humans saying it 'may have given' harmful information*. Fortune Well. https://fortune.com/well/2023/05/31/neda-ai-chatbot-harmful-advice/

Passantino, D. & O'Sullivan, J. (2023, May 22). *'Verified' Twitter accounts share fake image of 'explosion' near Pentagon, causing confusion*. CNN Business. https://www.cnn.com/2023/05/22/tech/twitter-fake-image-pentagon-explosion/index.html.

Xiang, C. (2023, May 25). *Eating Disorder Helpline Fires Staff, Transitions to Chabot After Unionization.* Tech by Vice. https://www.vice.com/en/article/n7ezkm/eating-disorder-helpline-fires-staff-transitions-to-chatbot-after-unionization.