Cognitive Hacking

How to Fight Fake News

Selena Groh
Computer System Security
Tufts University
December 13, 2017

Abstract

In a world where “fake news” consistently dominates the conversation and “post-truth” was Oxford Dictionaries Word of the Year 2016, the accuracy and authenticity of information is paramount. Yet attackers are finding more and more ways to produce and exploit misinformation. On an individual level, social engineering uses misinformation to trick single targets into revealing vital security information. On a larger scale, cognitive hacking weaponizes misinformation against the target through multiple users. Cognitive hacking is defined the practice of manipulating and falsifying information to induce changes in users’ perceptions. These changed perceptions then lead users to change their behaviors in ways that harm the target. For instance, a cognitive hacker might hack a Twitter account of a prominent newsperson, post a tweet reporting a scandal about a company, and watch the users who read the tweet injure the company by selling its stock and boycotting its products. Guarding against such information attacks is challenging as often only human research, fact-checking, and judgement can distinguish real from fabricated information. Thus, research into cognitive security measures such as information verification algorithms and collaborative filtering is exceedingly important. In this paper, I will present legislative, corporate, and personal methods for combating cognitive hacking and suggest potential areas for further research.

Introduction

In our current culture, our relationship to truth has become more malleable than ever before. Information which can be proven conclusively as fact does not seem to have much more impact on human behavior than moderately convincing or compelling falsehoods. In part, this is due to the lack of importance many place on fact checking or questioning presented information. In the computer science realm, attackers use misinformation to weaponize our behavior against intended targets in a practice called "cognitive hacking." Simply speaking, cognitive hackers tamper with information which leads to our behavior changing in ways that they desire.

Cognitive hacking is more formally defined by George Cybenko et al. as "gaining access to or breaking into a computer information system to modify certain user behaviors in a way that violates the integrity of the entire user information system" [1]. Such attacks consist of "manipulating perception and waiting for altered reality to produce actions that would complete the attack" [1].

There are three main types of cognitive hacking: misinformation, defacing, and spoofing [1]. The first, misinformation, is the most dangerous, as it is often covert. One prominent attack of this type is the "pump-and-dump" scheme, wherein the attacker presents false information (usually online) about a stock so that it will increase in value, thereby allowing the attacker to sell it at an inflated price before it decreases again, causing financial loss for stockholders [1]. This practice is illegal and investigated by the Securities and Exchange Commission (S.E.C.). However, it can be hard to fight as, according to Richard Walker, S.E.C. Director of Enforcement,

"on the Internet there is no clearly defined border between reliable and unreliable information" [2].

In contrast to misinformation, defacing is far more overt. Cognitive hackers defacing websites, social media accounts, or forums usually do so for attention, recognition, and satisfaction [1]. These defacements could consist of messages proclaiming their success in obtaining access to the target or content to which they want to call the attention [Figure 1].

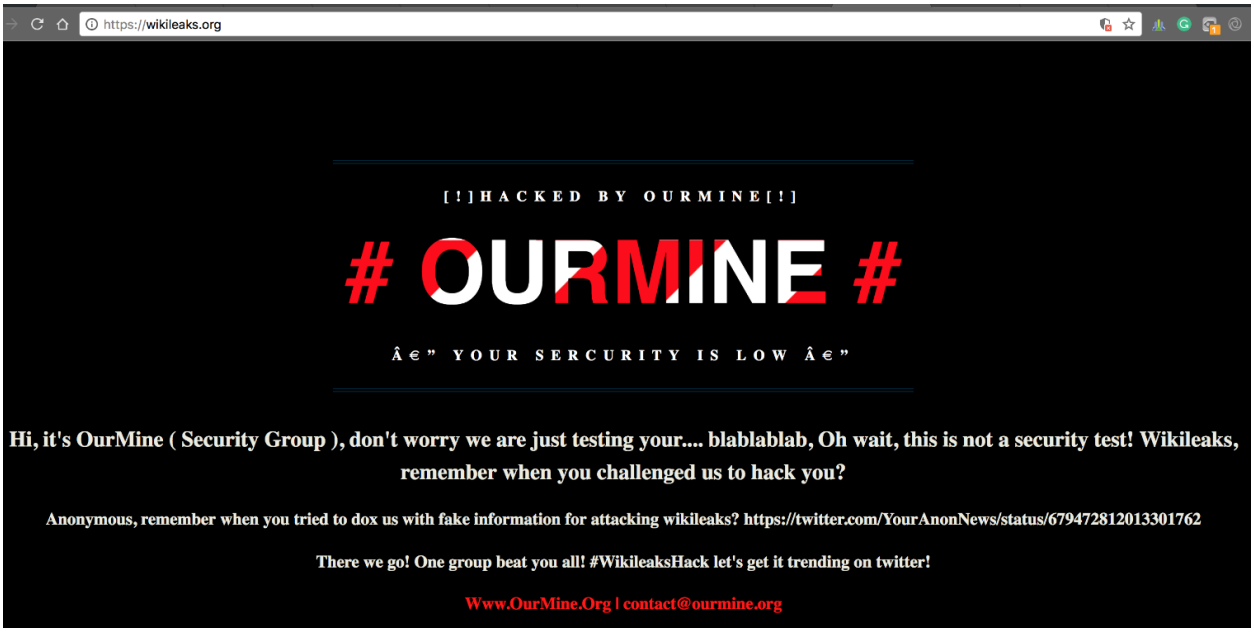

Figure 1. WikiLeaks.org defaced by the security group OurMine [3].

Such attacks, when less overt, pass into the realm of spoofing. Spoofing is where the attacker creates a fraudulent website and attempts to con the viewer into believing that the fraudulent website is the true one, all without tampering with the true site [Figure 2]. Since these attacks are covert, the attacker's motivations tend to be for exploitation. They may want to convince users to believe fake news (in which case this falls under misinformation as well) or to provide confidential information for use in later exploits.

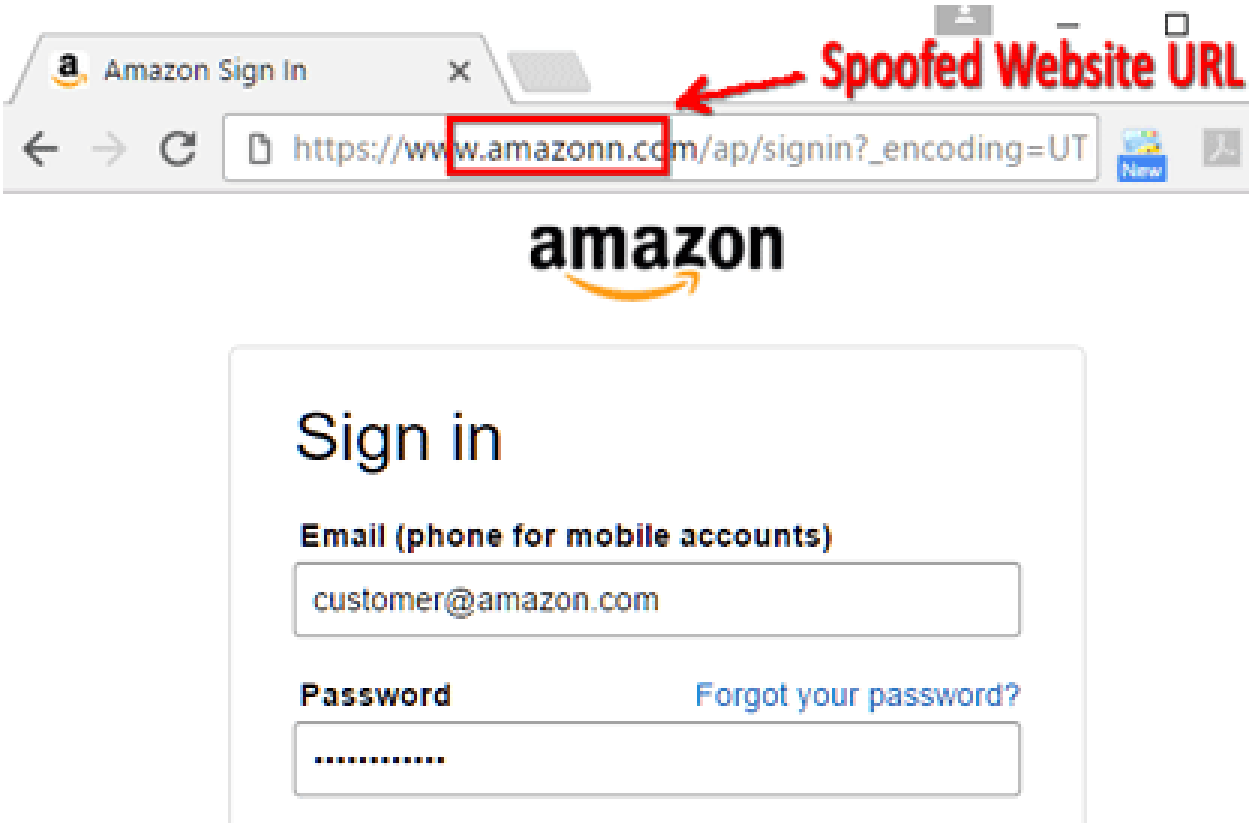

Figure 2. Spoofed Amazon Sign In Page [4].

These methods of cognitive hacking are strengthened by the ever-rising prevalence of the internet. We are increasingly reliant on the internet for dissemination and gathering of information. As such, it is paramount that we implement methods to guard against such attacks, as the information available on the internet is often used to make decisions with far-reaching consequences.

To the Community

Recently, the veracity of information has come to the forefront of our collective consciousness, with devastating consequences. One such example is the 2016 United States presidential election, wherein a preponderance of fake news [5] led many to question the results, thereby threatening the nature of our democracy itself. Russia's history of internet

trolling and misinformation campaigns are particularly relevant given recent investigations into their potential interference in the election [6]. While cognitive hacking did not account for all the fraudulent information available to the American public, it certainly provides a vehicle for malicious actors to manipulate events of national importance such as elections. The goal of this paper is to assert the threat cognitive hacking attacks pose and to suggest and encourage development of methods to combat them.

Defenses

There are three parties responsible for the identification and prevention of cognitive hacking: governments, companies, and users.

Governments

Governments' primary tool against cognitive hackers is legislation. While practices such as "pump-and-dump" schemes are illegal, often enforcement of such policies is inconsistent. For example, consider two case studies: NEI WebWorld and Jonathan Lebed. In late 1999, three young men acquired almost all of the bankrupt NEI WebWorld Corporation's shares at a very low price. They then posted on internet forums through multiple fake accounts to spread the rumor that a telecommunications company (also played by them) was acquiring NEI WebWorld. In the course of a single day, the price rose dramatically from $0.13 (the price at which they bought the shares) to over $15 [7]. The cognitive hackers realized a $364,000 profit from this 11500% increase in price, and they dumped their shares before the price dropped dramatically back to $0.25 [7]. The attackers were ordered to pay thousands of dollars in restitution and two were incarcerated.

However, Jonathan Lebed faced easier retribution. As a fifteen-year-old, he posted misleading forum messages under over 200 different names to encourage others to buy certain stocks. Over the course of six months, he gained $800,000 [1]. The S.E.C. caught on and he was initially ordered to give up all his profits. However, after appeal, he was allowed to keep some of his earnings, in part due to a lack of clarity on the legality of his actions [7]. While he did willfully mislead the public, one could argue that they acted of their own accord.

Such cases illustrate the importance of clarifying and strengthening legislation surrounding misinformation campaigns and cognitive hacking in general. Beyond more actively pursuing perpetrators of cognitive hacks, governments can also help fund research into cognitive hack identification algorithms. For instance, Gowri et al. proposed a method of identifying spoofed websites using elements of the URL domain name [8] and Taalohi et al. proposed a method for the same purpose using machine learning [9]. Such methods are promising and deserve further investigation, particularly given the significant consequences victims of phishing and cognitive hacking in general may face.

Companies

However, government legislation can only do so much to combat cognitive hacking; companies, particularly information channels such as news sites, must have robust measures for identifying and preventing such attacks. As of yet, there are not as many methods as desirable for identification of fraudulent information, primarily because misinformation that looks plausible to humans may often fool algorithms as well. However, that is not to say that there aren't some measures which can be of help. One method, suggested by George Cybenko, Annarita Giani, and Paul Thompson, is a modification of the Ulam games. It consists of checking

the veracity of assertions programmatically by confirming sub-assertions through independent sources [1]. Each piece of information would have a sequence of questions which the program would answer using external sources. If all the questions are satisfied correctly, the information is marked as valid.

While Ulam games rely on internal verification, collaborative filtering allows companies to utilize human judgement to identify fake news. Google has modified its search engine algorithm to incorporate user feedback and negatively weight pages deemed fraudulent [10, 11]. It also blocks such pages from using its AdSense network [10]. Similarly, Facebook aggregates reports of fraudulent accounts which malicious actors use to share fake or aggressive articles and comments [11]. Facebook has also begun implementing a feature which flags disputed articles with a warning for users [Figure 3]. However, Facebook's unwillingness to share data about the efficacy of this feature and lack of resources dedicated for it indicate that further work is necessary to truly combat misinformation on the social networking platform [12]. Future development might implement an icon similar to Twitter's "Verified Account" checkmark in order to mark information as independently substantiated.

Figure 3. Facebook's flag for disputed news [12].

While companies can certainly implement features to detect cognitive hacking, they can also do far more to prevent it in the first place. Website vulnerabilities, lack of authentication, and lack of independent verification all provide opportunities for cognitive hackers. An October 2001 attack on CNN.com is one such example that could have been prevented through more security-conscious web development. A cognitive hacker exploited a bug in CNN.com's "E-mail This to a Friend" button, eventually promoting a false story titled "Singer Britney Spears Killed in a Car Accident" to the top of the real site's "Most Popular Articles" page, where it was viewed more than 150,000 times [7, Figure 4]. If the vulnerability in the "E-mail This" button had not been present, the attacker would not have been able to gain such attention.

Figure 4. Spoofed CNN site with fake Britney Spears article [1].

In addition, companies must take care to maintain their web domain name registration. Clever phishers monitor expiring domain names to buy them should they become available. They can then use the domain name to their own ends, putting up a spoofed copy of the legitimate site or redirecting the user to any number of other potentially malicious applications. For example, scientific journals have recently gained attackers' attentions as many do not place as high a priority on their web maintenance as is necessary. As such, some attackers have been able to buy journals' web domains once they expire and replace the journals' content with their own spoofed versions [13]. They can then use these versions to phish email addresses, credit card information, and residential addresses. It is essential that companies realize the value of their domain name registration to prevent such opportunities.

Furthermore, companies can programmatically authenticate the author of information before they share it. Using certificates, news organizations may be able to verify that a press release came from a trusted source or than an article was truly posted by an organization. In

addition, linguistic analysis of multiple postings under different names may be able to identify a common author. Such a technique could be used by companies to identify multiple posts written by the same user, such as the Lebed attack, and warn users of potentially fraudulent behavior [1].

Finally, news organizations must independently verify stories before reporting on them themselves. In their eagerness to release information the fastest, news organizations often don't place as much of a priority on accuracy. However, such practices can have significant consequences. For example, consider an August 2000 cognitive hack on Emulex Corporation. 23-year-old Mark Jakob released a fake press release through Internet Wire Incorporated, an old employer, announcing that the S.E.C. was investigating Emulex [7]. This story was quickly picked up by legitimate financial websites and as such the company's stock dropped from $104 to $43 per share, earning Jakob $236,000 and costing the company $2.2 billion [1]. The key vulnerability in this hack is the legitimate websites' failure to identify the press release as fraudulent. They could have done so by contacting the corporation, the S.E.C., or any number of independent sources.

Users

If governments and companies fail to catch cognitive hacking, there are still some things users can do to help mitigate their susceptibility to such attacks. However, cognitive hackers have an advantage: internet communication makes users' judgements of tone, sincerity, and veracity difficult as we are unable to meet the source of the information face-to-face. Though content is important, biologically, humans tune into visual and auditory markers in body

language and tone of voice to determine if someone is telling the truth. Identifying misinformation online, then, must rely on different tactics.

When encountering an article with perhaps slightly far-fetched assertions, users should first review the contents of the article itself "beyond the headline," looking for inconsistencies or irregularities [14]. Users should then attempt to locate alternate and credible sources which report the same information [14]. If possible, they should attempt to locate the primary source of the information, or the earliest report they can find referencing the information. While a clever cognitive hack can evade such measures, these steps can help train users to question presented information.

Conclusion

Cognitive hacking at its core is an extension of the physical world of computer systems security into the behavioral world, where humans themselves become the tools of attackers. As such, it can be difficult to fight. As these attacks are designed to fool human beings, effective counter-measures often must involve mimicking of human cognition through complex algorithmic techniques. Yet, increased legislation, more secure applications, user feedback, independent verification, linguistic analysis, and increased user skepticism are all potentially powerful tools against misinformation, defacing, and spoofing. Future development in this field should explore more advanced techniques of mimicking human cognition through machine learning algorithms as well as more accurate automatic fact-checking systems. Given the potentially monumental consequences of cognitive hacks on financial, political, and personal levels, it is imperative that governments, companies, and users themselves place a high priority on thwarting such vulnerabilities.

References

[1] G. Cybenko, A. Giani and P. Thompson, "Cognitive hacking: A battle for the mind," *Computer,* vol. 35, no. 8, pp. 50-56, 2002.

[2] Bloomberg News, "The S.E.C. Accuses 23 of Internet Fraud," *The New York Times,* 2 March 2001.

[3] A. Hern, "WikiLeaks 'hacked' as OurMine group answers 'hack us' challenge," *The Guardian,* 31 August 2017.

[4] F. Stroud, "Webopedia," IT Business Edge, [Online]. Available: https://www.webopedia.com/TERM/W/website-spoofing.html. [Accessed 13 December 2017].

[5] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives,* vol. 31, no. 2, pp. 211-236, 2017.

[6] H. Berghel, "Oh, What a Tangled Web: Russian Hacking, Fake News, and the 2016 US Presidential Election," *Computer,* vol. 50, no. 9, pp. 87-91, 2017.

[7] G. Cybenko, A. Giani and P. Thompson, "Cognitive Hacking," in *Advances in Computers*, vol. 60, Elsevier, 2004, pp. 35-73.

[8] R. Gowri, V. K. Gandhi and M. Suriakala, "An efficient algorithm to identify phishing sites using URL domain features," *International Journal of Advanced Research in Computer Science,* vol. 8, no. 7, pp. 508-510, 2017.

[9] M. Taalohi, N. Langari and H. Tabatabaee, "Identifying phishing websites by techniques hyper heuristic and machine learning," *Science International,* vol. 27, no. 3, 2015.

[10] A. Chowdhry, "Facebook Launches A New Tool That Combats Fake News," *Forbes,* 5 March 2017.

[11] M.-L. Bârsan, "Military trolls, public distactions and the cyber," *Studia Universitatis Babes-Bolyai.Studia Europaea,* vol. 62, no. 2, pp. 17-29, 2017.

[12] E. Hunt, "'Disputed by multiple fact-checkers': Facebook rolls out new alert to combat fake news," *The Guardian,* 21 March 2017.

[13] J. Bohannon, "How to Hijack a Journal," *Science,* vol. 350, no. 6263, pp. 903-905, 20 November 2015.

[14] N. Jain, "How can we spot the fake news," *Express Computer,* 18 July 2017.