Kevin Chen, Will Berger, Emily Herron

Collaborative Project Report

Our project focused on the ethical implications of integrating AI in phishing attacks and prevention. We chose to use a poster as our central medium for the fair, and we also created handouts and a fake Vanderbilt Single Sign-on page to give an example of sophisticated phishing practices. We split our poster into 6 main components: a description of phishing, how AI is being used to protect us, how AI is being used against us, an "under-the-hood" of Vanderbilt's email quarantine issue, techniques to protect oneself from phishing, and a conclusion of the ethical implications. We chose a poster because it was likely the best way to present our findings to passersby in Sarratt. We could take attendees through a quick pitch of our poster, highlighting main points and offering them more in-depth information if they chose to read through the poster. After demonstrating our poster, we showed attendees a fake Vanderbilt Single Sign-on page which was hosted at the domain ssovanderbilt.com. The true sign-on page, of course, uses sso.vanderbilt.edu, but innocuous changes are one of the main tools of phishers. Lastly, we offered attendees handouts that served to educate them on phishing and ways to prevent it. So, they could keep a few tips and tricks at their desks before opening a potentially suspicious email.

Our main message for this project was three-fold: 1) it is much easier to get phished than you might imagine, 2) adversarial AI can be used to make phishing even more effective, and 3) we need to stay on the cutting edge of this application of AI to stay one step ahead of hackers. We drew upon two common discussion points in class for our ethical conclusion: 1) there exist sublated binaries between assistive and adversarial AI in these particular applications, causing a technological arms race already in motion, and 2) we must consider the caution with which we approach flagging potential false positives at the expense of creating additional harm from missed emails.

Companies like Google and Microsoft claims to use AI to prevent against phishing attacks that are otherwise nearly impossible to detect, with Google's Machine Learning tool, TensorFlow, blocking an additional 100 million spam emails each day. Google says they already block 99.9% of all spam, but catching the last sliver is only made possible through software that can learn and adapt to new patterns in spam emails [1]. Meanwhile, research organizations like ZeroFox have showed that machines are much better than humans at getting people to click on malicious links because they can create highly customized, innocuous emails [2].

We wanted to tie in an application of assistive AI that would be relevant to the many Vanderbilt students passing by our project, so we chose to include the story about Vanderbilt's email quarantine mistakenly locking away students' interview invitations last fall. One our the main objectives was to understand how Vanderbilt's email quarantine system works and how AI is used to classify phishing attacks from legitimate emails. So, we called VUIT and asked them about the issue and how it arose. Our main takeaway was that flagging false positives only becomes a problem if the system lacks transparency. The issue was not that interview invitations were being locked in a layer above spam, but that students were unaware that this was even happening. Our handouts also included a summary of the incident.

While actually phishing the class would have likely been an effective (though dangerous) experiment, having a demonstration of a fake Vanderbilt Single Sign-on page was also effective to demonstrate how simple and deceptive phishing can be. While we didn't use AI to build our demo, we asked attendees to imagine auto-generated log-in windows along with fraudulent emails from email addresses posing as friends or coworkers - all built by an AI. It is not a hard thing to imagine considering our review of AI chatbots and AI generated artwork in this class.

Furthermore, we really took away an appreciation for the relevance of this issue. These are not far-reaching, future applications of AI. These are issues that affect us right now. Email - an early application of the internet - has become something we use every day and take for granted; however, it is certainly still evolving. Failing to build sensible systems or considering every ethical and unethical use-case can result in important information getting lost, or worse - sensitive information being stolen.

Our group's experience at the class fair was overall extremely positive. Our poster effectively attracted attention, and attendees showed real curiosity and concern after we began explaining phishing and its dangers. People were genuinely engaged by the phishing material and looking 'under-the-hood' into Vanderbilt's email quarantine system. Simultaneously, we believe we did a good job keeping our explanation of AI and its implementation simple enough for others to understand. The live demonstration of the Vanderbilt Single Sign-on phishing website was very effective in conveying the difficulty of identifying real website from fake phishing websites. Our conversations with others helped us spread awareness of the serious ethical concerns in using AI for assistive or malicious purposes.

We had a very productive discussion with one particular attendee who gave us fantastic insight into how older, less technologically savvy individuals feel particularly vulnerable about being phished by malicious emails mimicking innocuous ones. The people we talked to over the course of the class fair helped us understand the lack of current knowledge within the general public about how easily and unknowingly they can be misled to voluntarily give up their personal information. We were commonly asked "do you think we will ever fall behind phishers using AI better than we can?" In short, probably not at a macro level. AI is generally very successful (if not overly cautious in the case of Vanderbilt's quarantine) when it comes to flagging potential attacks. As long as individuals use best practices when it comes to using the internet and visiting dangerous sites, we believe most people will be fine.

Lastly, the recommendations we gave to help people protect themselves from phishing attempts were greatly welcomed and appreciated by the people we talked with who were particularly curious about how to correctly identify common phishing attempts that mimic trustworthy email senders and website addresses. Most people left our display "scared but thankful." Most people thought of phishing as silly, half baked schemes that only careless or vulnerable people fall for. We left them with the idea that anyone can be the victim of a clever phishing attack, but that is knowledge that improve their ability to protect themselves. Our handouts that stressed how to protect oneself from phishing emails were quite popular and frequently requested by people we

spoke with. Overall, the class fair was very insightful into the real ethical impact and significance that our AI and phishing research has on the public.

Bibliography

[1] J. Vincent, "Gmail is now blocking 100 million extra spam messages every day with AI," *The Verge*, 06-Feb-2019. [Online]. Available: https://www.theverge.com/2019/2/6/18213453/gmail-tensorflow-machine-learning-spam-100-million. [Accessed: 14-Apr-2019].

[2] J. Seymour and P. Tully, "Automated E2E Spear Phishing on Twitter," *Black Hat*. [Online]. Available: https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter.pdf. [Accessed: 10-Apr-2019].