Automatic phishing detection versus user training, Is there a middle ground using XAI?

Sara Albakry[1,2] and Kami Vaniea[1]

[1] University of Edinburgh, Edinburgh, United Kingdom
[2] Umm Al-Qura University, Makkah, Saudi Arabia
{sara.albakry, kami.vaniea} @ed.ac.uk

Deciding when it is safe or unsafe to click on a received link is key to preventing phishing; where malicious actors deceive web users into providing access to their computers or providing confidential information. Despite improvements in the effectiveness and usability of anti-phishing solutions over more than two decades, phishers are succeeding in gaining the trust of vulnerable users at a higher rate than security systems are able to demonstrate the untrustworthiness of those malicious sites. Successful attacks can lead to serious consequences such as financial, data, and identity loss. For instance, the UK economy lost more than 280 million pounds in 2016 alone [1]. Today, we are combating phishing using two major approaches: automated detection and user training.

Automatically detecting and removing phishing communications from reaching the user's inbox is generally a good idea but it comes with few side effects. First, users have less opportunities to learn what phishing communications look like. Second, users start to develop high trust in communications reaching their inbox which is not always correct; especially with attacks such as spear phishing which only targets a small number of people [2]. While automated phishing detection systems continually strive for high accuracy and low false negatives and positives, it is unlikely that they will ever reach 100% accuracy [6]. These situations lead to the other approach: relying on users' judgment.

Upfront training [4], as one of many other user training solutions, provides the user with an opportunity to learn how to judge the trustworthiness/ untrustworthiness of communications such as websites or emails; however, it comes with few shortcomings. First, current training strategies require a considerable amount of user time and have shown inconsistent long-term effect on users' clicking behaviour. Additionally, many would argue that users are not even interested in investing their time in such training especially with their perception of being at risk being low [5]. Finally, judging the trustworthiness of communications is a difficult task for users because it requires uncommon skills such as reading a URL for the purpose of predicting its destination; which our research group has identified as a problematic point for end-users [3].

Towards supporting users' clicking decisions: We propose the design of a supportive environment, a middle ground between the two previous approaches. It could take the form of a communication system integrated into a web browser warning interface, which aims at improving users' comprehension of the warning situation and informing users' clicking decision, instead of them being confused and making an arbitrary clicking decision. We focus on situations where a user has clicked on a URL and the web browser has generated a warning stating that the website is suspicious; which may or may not be malicious. Designing such system entails the exploration of the following:

Can explainable AI help users understand the reason a site is classified untrustworthy? It seems natural to apply an explainable AI approach in this context to allow users to learn from the system for the four following reasons. First, recognizing phishing URLs is a critical security task to avoid malicious URLs. Second, automatic phishing detection systems are not enough. Third, explaining URLs to users is an effective strategy in changing users' clicking behaviour but has no lasting impact after training. Fourth, a phishing detection system is an approximate representation of security experts analysis that is not leveraged in user training.

Can Question-Answering Dialogue help users collaborate with AI systems when making evidence-based decisions? Fortunately, users have some skills and knowledge not readily available to the automatic phishing detectors. One of the important skills is a sense of contextual awareness. They know what bank they actually bank with, they also know about the current internal norms of an organization and how their co-workers typically write emails. Users ,therefore, have skills that they could use towards identifying phishing and helping the system generate more contextualized recommendations for verifying a site legitimacy.

Conclusion: Clicking on URLs is vital for both surfing the web and activating phishing attacks. Hence, web users strongly need a supportive environment to help them make informed clicking decisions on a case-by-case basis. However, designing a system that could achieve this goal in a scalable and usable way is challenging due to the complexity of URLs and users' context. In this position paper, we proposed the exploration of two AI approaches: XAI and question-answering dialogue. Those approaches could possibly help contextualize the feedback of phishing systems, but without further research, it is unclear how helpful they could be in guiding users' clicking decisions.

References

1. Annual Fraud Indicator 2016 http://www.port.ac.uk/media/contacts-and-departments/icjs/ccfs/Annual-Fraud-Indicator-2016.pdf
2. Aleroud, A., Zhou, L.: Phishing environments, techniques, and countermeasures: A survey. Computers & Security **68**, 160–196 (2017)
3. Althobaiti, K., Vaniea, K., Zheng, S.: Faheem: Explaining URLs to people using a Slack bot. In: Symposium on Digital Behaviour Intervention for Cyber Security (AISB) (april 2018)
4. Canova, G., Volkamer, M., Bergmann, C., Reinheimer, B.: NoPhish App Evaluation: Lab and Retention Study
5. Herley, C.: So long, and no thanks for the externalities: The rational rejection of security advice by users. In: Proceedings of the 2009 Workshop on New Security Paradigms Workshop. pp. 133–144. NSPW '09, ACM, New York, NY, USA (2009)
6. Zhang, Y., Egelman, S., Cranor, L., Hong, J.: Phinding phish: Evaluating anti-phishing tools. ISOC (2006)