Thesis Project Portfolio

Deep Learning Phishing URL Detection

(Technical Report)

Assessing Explainability in XAI

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Austin Huang
Spring, 2024
Department of Computer Science

Table of Contents

Sociotechnical Synthesis

Technology simplifies many of the interactions we have day to day in a way that can potentially be dangerous. For example, advertisements now are able to target user interests and hyperlinks allow users to easily navigate across the internet, but these links hidden inside an otherwise non-suspecting advertisement or email may end up being harmful phishing links, designed to cause damage. As another example, the recent advances in the field of artificial intelligence has made using AI tools for the non-technical user commonly accessible, with tools such as ChatGPT. These tools perform the creative work that allows users to skip many of the steps they themselves would have to take themselves from simple questions to even assistance in life-affecting scenarios. If errant or biased, these tools can misdirect the trust of users, who are merely looking for a way to streamline or make their work easier. At the intersection of these two examples, my thesis seeks to explore how an application of deep learning can help users avoid malicious phishing links and also how we keep the growing power that these technologies hold accountable with respect to user trust with the emergence of explainable AI.

For the technical side of my thesis, I write about a summer internship experience, where a team of interns and I participated in a hackathon to develop a basic phishing-link detector. Using Random Forests with Python's scikit-learn library combined with a Convolutional Neural Network and a Recurrent Neural Network, my team was able to build a model that when trained against a small training dataset, was able to perform with high accuracy to detect various phishing links in a small testing dataset. This summer project, while it was no breakthrough in development, was a useful introduction to the world of machine learning that supplemented my knowledge and allowed me to be proactive in finding an opportunity to apply my skill set.

In my STS research project, I investigated the developing field of explainable AI, asking the question of whether explainable AI is indeed successful in bridging the gap between the increasingly complex and data-backed decisions of AI and a common, less-knowledgeable user. To do this, I looked at the Defense Advanced Research Projects Agency's (DARPA) explainable AI research program, which occurred recently from 2015 to 2021. Within their program, DARPA funded various technical teams that explored different aspects of explainable AI. I selected three of these research-funded projects and looked at saliency maps, user interfaces, and frameworks of explainable AI and concurrent research happening outside of DARPA as resources for my research. Using Actor Network Theory, I examined the interactions between the developer, the user, and explainable AI. My research revealed that for the continuing effort to improve a model's explainability, developers may want to consider not merely how a user understands the explainable AI, but more importantly how an explainable AI understands its users. Additionally, it would be helpful to shift from the conventional understanding of evaluating a single answer for its effectiveness, and rather focus on the holistic dialogue between explainable AI and user to evaluate a model's suitability to adapt to a specific user's needs.

Through the research of both projects, I learned about the vulnerability of users when using various tools and how easy it is to gain and mishandle a user's trust. Perhaps with a simple phishing link detector, the bias of a model may not be as threatening as compared to a model that helps adjudicate legal cases or a model that helps with medical diagnosis. Regardless, with all these examples, it is apparent that AI can now be used to help users like no other computational technology before, and it emphasizes a need to develop a user-centric accountability system. This system, rather than continuing to focus on better quantitative metrics and algorithms, may be suited better to consider a psychological perspective and consideration of trust.