Explaining the Suspicion: Design of an XAI-based User-Focused Anti-Phishing Measure

Kilian Kluge
University of Ulm, Germany

Regina Eckhardt
University of Ulm, Germany

Explaining the Suspicion: Design of an XAI-Based User-Focused Anti-Phishing Measure

Kilian Kluge[1] and Regina Eckhardt[1]

[1] University of Ulm, Institute of Business Analytics, Ulm, Germany
{kilian.kluge,regina.eckhardt}@uni-ulm.de

Abstract. Phishing attacks are the primary cause of data and security breaches in businesses, public institutions, and private life. Due to inherent limitations and users' high susceptibility to increasingly sophisticated phishing attempts, existing anti-phishing measures cannot realize their full potential. Against this background, we utilize methods from the emerging research field of Explainable Artificial Intelligence (XAI) for the design of a user-focused anti-phishing measure. By leveraging the power of state-of-the-art phishing detectors, our approach uncovers the words and phrases in an e-mail most relevant for identifying phishing attempts. We empirically show that our approach reliably extracts segments of text considered relevant for the discrimination between genuine and phishing e-mails. Our work opens up novel prospects for phishing prevention and demonstrates the tremendous potential of XAI methods beyond applications in AI.

Keywords: Phishing Prevention, Explainable Artificial Intelligence, Interpretable Machine Learning, User-Centric XAI

1 Introduction

During the first weeks of the COVID-19 pandemic, a large number of US citizens received an e-mail ostensibly from their employers' payroll department. The e-mail informed them that the federal government was considering a financial relief package, entitling them to a $1000 check. In order to benefit from this measure, they would need to verify "their email account for new payroll directory" by following a "Secure Link" included in the e-mail [1]. These e-mails are exemplary for a phishing attack: To gain sensitive information for malicious purposes, the sender imitates a trustworthy source and promises a personal benefit to deceive the user [2, 3].

Phishing attacks are the primary way in which identity theft and security breaches occur in businesses, public institutions, and private life [4–6]. Virtually all users of electronic communication are frequently subject to phishing attempts [6–8]. In light of this perpetually growing threat, IT security researchers and practitioners have developed a large variety of anti-phishing measures. Commonly, these are divided into three categories: Blocking malicious e-mails before they reach users, warning users, and training users not to fall for phishing [2, 4, 6, 8]. All of these measures are applied

16th International Conference on Wirtschaftsinformatik,
March 2021, Essen, Germany

in practice with some success. Ultimately, however, due to inherent limitations, neither is effective in preventing phishing attempts from succeeding. While state-of-the-art phishing detectors that aim to identify and filter out phishing attempts are robust and versatile, they suffer from their limited accuracy [2, 8, 9]. In order to avoid a high number of false positives – i.e., mistakenly blocking genuine e-mails – the detectors are generally tuned for maximum precision [6]. Consequently, many e-mails that a detector identified as suspicious of constituting a phishing attack reach the user [2, 8, 9]. Therefore, user behavior is of paramount importance in phishing prevention.

While the response rate to phishing e-mails varies widely between users and particular variants of phishing attacks, on average, 10% to 20% of users that receive a phishing e-mail act on it [5, 10, 11]. Anti-phishing training aims to reduce this rate by educating users on how to identify phishing attempts [2, 8]. However, while users successfully learn to spot telltale signs of phishing, they nevertheless fall for it in everyday situations, which is overwhelmingly attributed to a lack of awareness when performing routine tasks in a familiar and trusted environment [10–16]. Existing anti-phishing measures that aim to raise the users' attention, such as warning messages, are often ignored, as they are perceived as too generic [4, 6–8]. In summary, on the one hand, the need to avoid false positives prevents phishing detectors from realizing their full potential [6]. On the other hand, users do not benefit from the knowledge gained in anti-phishing training and remain susceptible to phishing because, in everyday life, they lack the required attention [11, 12].

Against this background, approaches from the emerging field of Explainable Artificial Intelligence (XAI) [17, 18] harbor to date untapped potential for the design of more effective user-focused anti-phishing measures [9]. In particular, XAI methods designed to explain the classification of text documents by black-box models [19, 20] could convey to the user which elements in an e-mail most strongly influenced a phishing detector in identifying it as suspicious. These explanations, which could be provided for all e-mails that a detector had to let pass to avoid false positives, constitute highly specific warnings that are expected to effectively raise the users' attention [3, 9, 11, 14]. Pursuing this basic idea, we design a novel approach that identifies phishing cues in suspicious e-mails by generating explanations for the output of a phishing detector. Thereby, we not only pave the way for more effective user-focused anti-phishing measures but provide a glimpse of the potential applications of XAI in the realm of IT security and beyond.

Following the Design Science methodology [21], the remainder of the paper is structured as follows: In Section 2, we survey research from the fields of phishing prevention as well as XAI and conclude with the research gap. Subsequently, in Section 3, we describe the design of a novel XAI approach to extract cues and phrases from e-mails that contribute to their assessment by a black-box phishing detector. In Section 4, we demonstrate and evaluate the applicability of the approach using a real-world dataset. Subsequently, in Section 5, we summarize our findings and conclude our paper with a discussion of the limitations of our research and an outlook on future work.

2 Related Work and Research Gap

To lay the foundation for the design of our novel approach, in the following, we first summarize research on phishing attacks and the cues based on which both automatic detectors and users can distinguish phishing attempts from legitimate communication. Then, we provide a brief overview of existing anti-phishing measures and their respective strengths and drawbacks. Last, we introduce the research field of Explainable Artificial Intelligence (XAI) and survey XAI methods for explaining document classification.

2.1 Phishing Attacks and Phishing Cues

Phishing is a social engineering attack that aims to exploit specific weaknesses of users [4, 5, 7]. The attacker imitates a trustworthy source to gain sensitive information for malicious purposes [2, 4, 5]. A phishing attack typically consists of three phases [5, 6]: Circumventing IT security measures (e.g., a phishing detector) to deliver an electronic communication (e.g., an e-mail) to a user, convincing the user to engage in the intended activity (e.g., click on a link to a counterfeit website and enter their credit card details), and finally gaining from the attack (e.g., receive a payment). Most phishing attacks are carried out via e-mail and traditionally target a broad audience, e.g., all users of a popular online platform [5–7]. The e-mails include a link to a forged website, where users are asked to enter their login credentials, on which the attackers then capitalize. Increasingly, personalized attacks target employees of specific company departments or public offices using elaborately crafted e-mails that imitate communication by superiors or co-workers [5–7, 10]. Often, the goal is to initiate large payments or gain access to confidential information [5, 8, 10].

Researchers have identified cues that are helpful to distinguish between genuine and phishing e-mails through user studies [10, 13, 14, 22] and analyzing e-mails [3, 14, 23]. Among the main discriminatory elements are the sender's address and other technical information in the e-mails' header, the links included in the e-mail, and words and phrases in the e-mails' text [3, 10, 14]. In contrast, the graphical design of an e-mail, visual elements, and the presence of legal information (e.g., a disclaimer) are of little informative value [3, 14].

Textual information is arguably the most relevant for users when distinguishing between genuine and phishing e-mails. On the one hand, increasingly sophisticated imitation of the style and design of e-mails renders these features unsuitable as discriminators [11, 14]. On the other hand, textual cues such as urgency require complex judgment and background knowledge [10, 23]. Thus, in contrast to technical cues (e.g., URL spoofing), they often cannot be unambiguously detected by automated filters [5–7]. Indeed, anti-phishing training places emphasis on textual cues and caution users' against just considering the superficial properties of an e-mail [3, 11]. Table 1 summarizes typical categories of textual phishing cues.

Table 1. Typical categories of textual cues in phishing e-mails [2, 3, 10, 11, 14, 22, 23]

Category	Example from the IWSPA v2.0 dataset [24]
Urgency	You have 72 hours to verify the information, …
(Appeal to) authority	A message from the CEO …
Importance	We have reason to believe that your account was accessed by a third party.
Positive consequence (reward)	In return we will deposit $70 to your account …
Negative consequence (loss)	If you do not verify yourself, your account will be suspended.
References to security and safety	Security is one of our top goals at our company …
Spelling mistakes and grammatical errors	You were qualified to participate in $50.00 reward surwey.
Lack of personalization	Dear Valued Customer, …

2.2 Technical and User-Focused Anti-Phishing Measures

Measures for phishing prevention are commonly divided into technical and user-focused anti-phishing measures. While the former aim to block malicious e-mails before they reach users, the latter intend to prevent users from falling for phishing attempts [2, 4, 6].

Technical anti-phishing measures detect phishing e-mails by searching for common characteristics [6, 7]. Typical approaches include rule-based filters and machine-learning-based detectors [6]. Filters are based on manually assembled blacklists [2, 25] and are thus inherently constrained to already known cues and patterns [5–7]. In contrast, machine-learning-based detectors learn to detect phishing e-mails from training on examples [6]. While earlier approaches relied on predefined features [7], modern deep learning methods autonomously identify intricate patterns in raw data and have demonstrated excellent performance in phishing detection [8, 26]. However, phishing detectors have to be configured such that no genuine e-mail is mistakenly classified as a phishing attempt and thus discarded [6, 9]. Indeed, it is the "concern over liability for false positives [that] is the major barrier to deploying more aggressive heuristics" [6, p. 79], which in turn limits the effectiveness of phishing detectors.

Depending on the type of attack and target audience [cf. 5], studies found that between 5% and close to 50% of users that receive a phishing e-mail fall for the attempt [5, 10, 11]. Against this background, user-focused anti-phishing measures aim to reduce users' susceptibility to phishing attacks. They comprise anti-phishing training as well as preventive mechanisms and warning facilities [2, 6]. Trainings aim to raise users' awareness of the threat and educate them on how to identify phishing attempts. They are administered in the form of resources for self-study (e.g., texts [8], videos [27], or games [28, 29]), classroom-style training, and interventional training [2, 8, 10, 11]. In the course of the latter, imitated phishing e-mails are sent to users. When they fall for the simulated attack (e.g., by clicking on an included link), they are immediately presented with self-study material [2, 6, 10]. However, anti-phishing training is not

sufficient to prevent users from falling for phishing attacks [2, 12]. While trainings have been shown to increase users' ability to identify phishing attempts when tasked to do so [2, 6], trained users nevertheless fall for phishing in everyday situations [8].

Researchers have theorized and demonstrated that the cause for users' high susceptibility to phishing is their lack of attention when performing routine tasks in a familiar and trusted environment [8, 12, 16, 30]. It is further amplified by users' tendency to underestimate their vulnerability to phishing attacks [10, 28, 31]. Thus, preventive mechanisms such as regular reminders [10], warning messages [4, 10], or tooltips that help users to evaluate URLs [32] are employed to motivate users to stay alert and scrutinize all communication for phishing cues [6, 10, 32]. However, users often overlook or outright ignore these warnings when they are passive indicators or not perceived as specific and relevant to their current situation [2, 4, 7, 8, 10].

2.3 Explainable Artificial Intelligence and Generation of Explanations for Document Classification

Since at least the rise of deep learning, AI systems have become ubiquitous. Thus, an increasing number of people are faced with the consequences of decisions and recommendations generated by effectively black-box systems [17, 33]. Against this background, the research field of Explainable Artificial Intelligence (XAI) focuses on automatically generating explanations for AI decisions [17, 18, 33, 34].

XAI methods can be distinguished by their aim and their dependency on a particular kind of machine learning model [18, 34]. In the context of explanations for AI systems for text and document classification (such as phishing detectors), both researchers and practitioners have taken a particular interest in outcome explanations [20, 35]. This kind of explanation is not concerned with revealing the inner workings of the AI system but aims to provide a human-understandable reasoning for one specific decision [34, 36].

One avenue to explain an AI system's decisions in this manner is through local feature importance [18, 34]. The underlying idea is to assign a weight to each of the input's features that reflects how strongly it contributes – positively or negatively – to the AI system's decision. The SHAP family constitutes a popular example of such methods [37]. Some of its variants are model-agnostic, i.e., do not require access to the AI system's internals and are thus applicable to any kind of AI system [36, 37]. A study by Weerts et al. [38] suggests that SHAP explanations succeed in drawing user's attention to particularly influential features that they would otherwise have overlooked. However, explanations based on local feature importance do not necessarily transfer to other decisions by the same AI system [19, 34].

This limitation is addressed by several more robust XAI methods, which can be divided into search-based approaches and document classifiers with integrated explanation capabilities. Martens and Provost [20] define "explanations" as minimal sets of words that, if removed from the particular document under investigation, change the classifier's prediction. To find explanations, they utilize a best-first heuristic search with search tree pruning. In the case of a non-linear classifier, two post-processing optimizations aim to ensure that the found set is indeed minimal [20]. Fernandez et al. [39] generalize this approach to replacing words instead of removing them and

introduce a variable cost for replacement, allowing for more fine-grained control of the explanations' properties. Similar to these "explanations," the "anchors" introduced by Ribeiro et al. [19] are sets of words. However, instead of constituting a minimal set of words required for the classification, "anchors" aim to be representative of the AI system. They are defined as a set of words that, if present, is sufficient to guarantee the classification independent of changes to the remainder of the document. "Anchors" are built up word by word through local beam search [19].

Instead of generating explanations post-hoc [34], Lei et al. [35] train two joint machine-learning models to find explanations for the classification of texts. While an "encoder" model classifies a text, a "generator" model extracts the corresponding "rationales," which are short phrases that, individually, are classified similarly as the full text. An objective function ensures both correct classification and the "rationales"' characteristics, namely conciseness and coherence [35, 40]. With their τ-SS3 classifier, Burdisso et al. [41] again pursue a different approach. τ-SS3 is inherently interpretable, i.e., the AI system itself transparently reveals which word sequences in a text stream contributed most to its output.

2.4 Research Gap

Phishing is a pervasive threat for businesses, public institutions, and private individuals alike. Technical anti-phishing measures filter out malicious e-mails with increasing effectiveness. However, due to their limited accuracy, phishing e-mails nevertheless reach the inboxes of users, which consequently have a decisive role to play [2, 6, 7, 9]. Despite efforts to educate users, they frequently fall for phishing attempts, in particular for those that are sophisticated imitations of genuine e-mails [8, 10]. It is, however, generally not a lack of knowledge or awareness of the grave consequences but a lack of attention in everyday situations that makes users vulnerable [10, 12, 16]. Existing preventive mechanisms such as warning messages often remain without effect, as users perceive them as too unspecific and disregard them [4, 7, 8, 10].

In light of the power of modern phishing detectors, methods from the field of XAI appear as a promising foundation for the design of more specific, and thus, more effective user-focused anti-phishing measures [9]. Following this idea, based on outcome explanation methods for document classification [19, 20, 35], we design a novel approach that uncovers words and phrases in e-mails that are telltale signs of phishing. Our work paves the way for user-focused anti-phishing measures that effectively raise users' attention and guide their assessment of suspicious communication [11, 15, 30, 31]. It further serves as an example of the potential of XAI methods to address problems of high practical relevance beyond the field of artificial intelligence.

3 A Novel XAI Approach to Uncover Phishing Cues in E-Mails

We design a novel XAI approach to draw the user's attention to the telltale signs of phishing in a suspicious e-mail. The underlying basic idea is to generate explanations for a phishing detector's assessment of an e-mail that serve as highly specific warnings.

The starting point for our approach is a phishing detector. In the following, we describe it as a model m that takes an e-mail x as its input and outputs a score $s \in [0, 1]$ and treat it as a black box otherwise. All incoming e-mails for which $m(x) = s > t_{phish}$ are considered phishing e-mails and are filtered out before they reach a user's inbox. Since the detection threshold t_{phish} has to be set such that no genuine e-mails are discarded [cf. 6], many e-mails to which the detector assigns a high score – and thus, a high likelihood of being a phishing attempt – nevertheless reach the user [6, 8, 9].

Three design decisions characterize our approach. First, to be widely applicable and to not adversely interfere with the phishing detector's performance, we design the approach to be model-agnostic [19, 39]. Second, we focus exclusively on textual cues, as these are most relevant to distinguish phishing from genuine e-mails and easiest to assess for laypeople [3, 11, 14, 23]. Third, to assist the users' assessment, we strive to highlight precisely the telltale signs of phishing (cf. Table 1 and Figure 1). For this, we identify the words and phrases in an e-mail that significantly contribute to the phishing detectors score. In the following, we describe the design of our approach in detail and elaborate on the design decisions.

3.1 Designing Explanations as Text Highlights

The goal of our approach is to assist users in reliably identifying phishing e-mails. Thus, the explanations produced by our approach should match how people evaluate e-mails [11, 33, 42, 43]. Phishing research suggests that textual cues are most relevant to distinguish between genuine and phishing e-mails (cf. Section 2.1). On the one hand, textual cues are easiest to comprehend and evaluate for laypeople [3, 14]. On the other hand, they are the only cues present in types of phishing e-mails that do not rely on technical manipulation [6, 10, 11].

Against this background, we design our approach to produce explanations in the shape of text highlights (cf. Figure 1). Specifically, we highlight short sequences of text [35, 41], which offers three advantages. First, people are familiar with this concept from everyday life [cf. 42]. Second, the interpretation of the explanations does not require technical knowledge about their production [19, 44]. Further, the focus on textual cues avoids the adverse effects of cognitive biases associated with quantitative indicators such as confidence scores [33]. Third, the interpretation of text highlights demands substantial cognitive effort and thus encourages thorough evaluation [42], which is favorable for users' ability to accurately identify phishing attempts [11, 15, 31].

To formalize the notion of text highlights, we represent an e-mail as a sequence of words $x = [x_0, x_1, \ldots, x_N]$ [35, 45]. A text highlight explanation can then be represented by a binary vector a of the same length as x, where $a_i = 1$ indicates that the word x_i is highlighted and $a_i = 0$ indicates that it is not.

3.2 Characteristics of Suitable Explanations

The basic idea of our approach is to convey to the user which words and phrases in an e-mail influenced a phishing detector's classification of the e-mail as suspicious. In the realm of XAI, the task of explaining a model's output by uncovering which parts of the input contributed to its assessment has attracted considerable research attention (cf. Section 2.3). In the following, we draw from this prior work to derive and define the characteristics of explanations required in our application context.

As worked out in the previous section, our explanations take the shape of text highlights. To ensure that the highlighted phrases indeed represent phishing cues, we demand that the phishing detector classifies them as suspicious themselves. In that regard, the explanations generated by our approach are similar to the "rationales" proposed by Lei et al. [35]. Taking into account that this assessment might be coincidental, we require the phrases themselves to be sufficient for the classification of the entire e-mail. More specifically, similar to the anchors defined by Ribeiro et al. [19], replacing the remainder of the e-mail with different words should have a negligible influence on the phishing detector's assessment [cf. 39].

We capture these characteristics in the concept of a document anchor. For its formal definition, we resort to the perturbation set D_x introduced by Ribeiro et al. [19]. For a given e-mail x, this set contains all possible variants z that can be generated by replacing words in x with either blanks or similar words [19, 39]. A particular sequence of highlighted words in an e-mail is a document anchor if it is present in most $z \in D_x$ that are classified similarly as the original e-mail, but not present in the $z \in D_x$ for which this is not the case. More formally, a text highlight described by a binary vector a is a document anchor for x if for any $z \in D_x$

$$|z \odot a| = |a| \Longrightarrow m(z) \geq m(x) - \tau, \tag{1}$$

where τ is an application-specific constant.

In general, many document anchors exist for any given e-mail x. However, not all of them constitute a good explanation [19, 42]. On the one hand, an anchor that covers the entire document ($a_i = 1 \; \forall \; i$) always fulfills the definition, but conveys no information to the user that is particularly helpful in distinguishing between phishing and genuine e-mails. On the other hand, while a few specific words might be sufficient to guarantee the correct classification, the user perceives text in phrases [41]. Thus, while prior work strives to find a minimal number of words in an explanation [19, 20, 39], the shortest possible explanation is not necessarily the best in the eyes of the user [40, 42]. Based on these considerations, we require that the document anchors chosen as explanations both contain an appropriate number of words and consist of at most a few connected phrases. We encode these characteristics in an objective function that takes on a minimum value for an optimal anchor:

$$\mathcal{O}(\mathrm{a}) = (|a| - l)^2 + \beta \cdot \textstyle\sum_i |a_i - a_{i-1}| \tag{2}$$

The first term measures how far the number of highlighted words contained in the document anchor described by a deviates from the desired target l. The second term measures the coherence, i.e., the number of connected sequences of words [35]. The

coefficient β weights the two terms and allows for fine-tuning of the explanations' characteristics.

3.3 Model-Agnostic Generation of Explanations for Suspicious E-mails

Up to this point, we have defined the shape of the explanations and developed the concept of document anchors to capture their desired characteristics. What remains in the design of our approach is to devise a method that, for a given e-mail x, generates a document anchor a that minimizes the objective function $\mathcal{O}(a)$ [cf. 19].

As our approach is based on an existing phishing detector, the search for a suitable anchor cannot make any assumptions regarding the model's inner workings. Therefore, we design our approach to be model-agnostic. This not only allows it to be used with any kind of phishing detector [18, 36]. It further ensures that the phishing detector's functionality and performance are not affected in any way [19, 40]. Conversely, the phishing detector's properties do not impose restrictions on the design of the method for the generation of explanations [34, 36].

Incorporating these benefits, we follow the general idea of search-based approaches [cf. 45]. The basic concept is to find and construct an anchor for an e-mail x by probing the detector with perturbed versions of that e-mail [19, 20, 39]. Addressing the requirement that our approach should generate explanations that consist of phrases, we construct an anchor a by combining individual phrases p ($a = \sum p$).

In our approach, we generate perturbed versions $z \in D_x$ by replacing words in the e-mail [19, 39]. In line with the definition of a document anchor, we iteratively search for phrases p that are present in those versions z that the detector identifies as suspicious, but absent from versions of the e-mail that the detector considers genuine. To this end, we utilize local beam search [19, 45], which we initialize with N seed phrases. Each iteration of the search consists of three steps. First, we generate N_{child} child phrases from each of the N phrases by growing, shrinking, or shifting the highlighted sequences of words. Second, we use the KL-LUCB algorithm [46] to determine the N best phrases among the $N \cdot N_{child}$ children [19]. For this, we estimate the expectation value for a $z \in D_x$ that contains the phrase p to be classified as suspicious by the model [19]:

$$\mathbb{E}(p) = \mathbb{E}_{|z \odot p| = |p|}[m(z) \geq m(x) - \tau] \tag{3}$$

We repeatedly refine these estimates until the lower bound on the expectation value of the N^{th}-best phrase surpasses the upper bound on the next-best phrase's expectation value by at least Δ_{min}. The N best phrases then form the set of N phrases for the next iteration. To boost convergence, we keep a set of the N_{elite} best phrases that we add to the child phrases in every round of the search [45]. In the third and final step of each iteration, we merge the current set of N phrases to an anchor candidate. If the objective function's value for this candidate falls below a previously specified threshold or the number of iterations surpasses a given maximum, the search terminates. Both the threshold and the maximum number of iterations, as well as the beam search parameters N, N_{child}, and N_{elite} influence the efficiency of the search and the consistency of the document anchors' characteristics [19, 45].

4 Demonstration and Evaluation

In the following, as an essential part of the Design Science research process [21], we demonstrate and evaluate the efficacy of our approach. For this, we instantiate it using a real-world dataset and conduct a series of summative evaluations adhering to the Framework for Evaluation in Design Science Research (FEDS) [47].

4.1 Dataset and Phishing Detector

The instantiation and subsequent evaluation of our approach requires a phishing detector and a set of both phishing and genuine e-mails. We use the English-language IWSPA-AP v2.0 dataset [24, 26] that was compiled to enable the comparison of machine-learning-based phishing detectors. It consists of 452 phishing and 3505 legitimate e-mails. We randomly select 80% of each kind for the training set and leave the remaining e-mails as the test set.

Using the training set, we instantiate a bidirectional LSTM (long short-term memory) recurrent neural network as the phishing detector, which is a standard model for text classification [45]. In line with real-world requirements [6], we aim to set the threshold above which we discard an e-mail as phishing t_{phish} such that the false positive rate is minimal. To avoid fatigue due to frequent unsubstantiated warnings, the threshold t_{susp} above which an e-mail is considered suspicious should be set such that the probability that these e-mails are indeed phishing attempts is reasonably high [6, 10]. We find that for the given detector and dataset, $t_{phish} = 0.98$ and $t_{susp} = 0.20$ achieve these goals, resulting in a false positive rate of 0.43% and the classification of 16 genuine and 11 phishing e-mails as suspicious. Just 2.2% of phishing e-mails reach the user without explanations.

4.2 Instantiation

Our approach generates text highlight explanations by performing a local beam search guided by an objective function and repeated estimation of the expectation value $\mathbb{E}(p)$ (Eq. 3). Accordingly, in the following, we parametrize the required components.

To generate the samples $z \in D_x$ needed to estimate $\mathbb{E}(p)$, we randomly replace words in the e-mail x with blanks. Since evaluating $\mathbb{E}(p)$ for a given phrase p requires a z for which $|z \odot p| = |z|$ (cf. Eq. 3), we can optimize the search's efficiency by maximizing the likelihood that this condition is fulfilled. As p generally consists of connected sequences of words, we do not randomly replace words but generate $z \in D_x$ that each contain a single sequence of varying length. To obtain an unbiased estimate of $\mathbb{E}(p)$, the unconditional probability $P(m(z) \geq m(x) - \tau)$ should be close to 0.5. We find that for the given phishing detector, $\tau = 0.15\, m(x)$ is a suitable choice. We generate at most 1024 samples $z \in D_x$ to limit the load on the phishing detector.

To instantiate the search component, we first parametrize the local beam search. We use a beamwidth of $N = 10$ and maintain an elite set of size $N_{elite} = 4$. We initialize the search with randomly placed phrases of three words. In each round, we generate

$N_{child} = 2$ new phrases from each of the N current best phrases by appending one word or shifting them in either direction. Finally, we parametrize the objective function (Eq. 2) with a target length of $l = 10$ and $\beta = 4$, which we find to strike a suitable balance between highlighting relevant phishing cues and comprehensibility. We stop when $\mathcal{O}(a) \leq 16.0$ or five iterations have passed. Figure 1 displays an example of an explanation generated by our approach.

… login to your account and give us the necessary information. Complete the necessary verification tasks within 5 days, or your account might get temporarily suspended. Proceed with the link below.	For more information on protecting yourself from fraud, please review our Security Tips. Protect Your Password: You should never give your PayPal password to anyone, including PayPal employees.

Figure 1. Example of text highlights generated by our approach for a phishing e-mail that seeks to persuade users to provide their PayPal login credentials by invoking a sense of urgency, suggesting impending negative consequences, and alluding to standard security practices.

4.3 Evaluation

As suggested by FEDS, we explicate the goals and evaluation strategy before designing particular evaluation episodes [47]. The goal of the evaluation is to investigate whether our approach succeeds in generating explanations for suspected phishing attempts that help users distinguish between genuine and phishing e-mails. Owing to our research's exploratory nature, the main risks in the design of our approach are technically-oriented. Thus, FEDS' "Technical Risk & Efficacy" strategy, which prescribes a series of increasingly summative and naturalistic evaluations, is an appropriate choice [47].

For the individual evaluation episodes, we utilize the established concept of functionally-grounded evaluation of explainable systems defined by Doshi-Velez and Kim [48] and assess explanations using three proxy measures. Each proxy measure operationalizes a particular goal of our design.

First, the highlighted segments of text should be classified similarly to the entire e-mail, i.e., as suspicious. Thus, we take the score that the phishing detector attributes to the text highlights as the corresponding proxy measure (*Score*).

Second, the explanations should be comprehensible for laypeople. For this, an explanation should consist of connected phrases rather than individual words scattered across the e-mail. Therefore, we take the number of highlighted sequences as the corresponding proxy measure (*Comprehensibility*).

Finally, to draw the users' attention to those elements in a suspicious e-mail relevant to assessing the threat, the highlighted parts of the text should represent phishing cues. To evaluate this, we let two researchers code the words in each of the suspicious e-mails according to the categories in Table 1 and measure the text highlights' overlap with the humans' assessment. To account for the vastly different amount of phishing cues in the e-mails (ranging from 0% to 50% of words), we divide this value by the ratio of cues expected to be found when randomly selecting words to be highlighted (*Relevance*).

To benchmark the values obtained for the proxies, we utilize two competing approaches: As the baseline, we create explanations by randomly highlighting $l = 10$ words in an e-mail (Random). Further, to assess the effect of the information our approach obtains from the phishing detector, we perform the local beam search with a fixed $\mathbb{E}(p) = 1$ (Search-only). To obtain statistically sound conclusions, we apply each approach fifty times for each of the 27 suspicious e-mails, assess the resulting explanations, and aggregate the results (Figure 2).

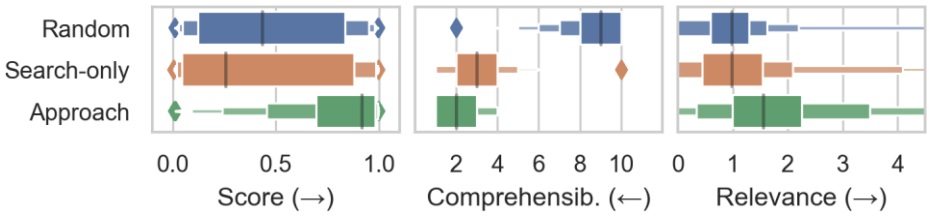

Figure 2. Aggregated evaluation results. The arrows indicate the direction of better values.

We find that our approach outperforms the competing approaches for all three proxy measures. First, the *Score* of the text highlights generated by our approach is significantly higher and exhibits a smaller variance ($1^{st}/2^{nd}/3^{rd}$ quartile .70/.92/.98) compared to Random (.12/.43/.84) and Search-only (.05/.29/.89). Second, our approach selects only 2.1 ± 1.0 phrases in an e-mail to be highlighted, rendering its explanations comprehensible. Third, despite selecting the fewest phrases, the words highlighted by our approach exhibit higher *Relevance* for distinguishing between phishing and genuine e-mails (.99/1.6/2.3) than the text highlights generated by Search-only (.44/.98/1.5), whose *Relevance* is similar to that of the Random baseline (.60/.98/1.3). The difference in *Relevance* is significant (Mann-Whitney $U = 1.20 \cdot 10^6$, $n_1 = n_2 = 1350$, $p < 10^{-3}$ one-sided, effect size 0.66), which validates that through $\mathbb{E}(p)$ our approach indeed extracts the required information on phishing cues from the detector.

In summary, our approach successfully generates explanations in the shape of text highlights that are well suited to draw the users' attention to phishing cues in an e-mail.

5 Conclusion, Limitations, and Outlook on Further Research

Phishing is a threat to businesses, public institutions, and private individuals alike. Current anti-phishing measures ultimately fail at effectively preventing users from falling for phishing attacks. Against this background, XAI methods offer a promising path towards more effective user-focused anti-phishing measures that leverage the power of state-of-the-art phishing detectors. Pursuing this idea, we designed a novel XAI approach that identifies telltale signs of phishing in suspicious e-mails. Building on research in phishing susceptibility and anti-phishing training, we designed its explanations to raise users' attention and assist their assessment of the potential threat. We demonstrated our approach utilizing a real-world dataset and a deep learning phishing detector. Rigorous functionally-grounded evaluation indicates that our

approach succeeds in producing explanations that are both relevant and comprehensible. In addition to the design of a novel XAI approach, our research contributes to theory and practice in two ways. On the one hand, it validates the feasibility of utilizing XAI methods for the design of user-focused anti-phishing measures. On the other hand, it serves as an example of how XAI methods can be applied to address problems of high practical relevance beyond the field of AI.

Although our work constitutes a substantial step, it is subject to several limitations that call for further research. First, by design, our approach can only uncover cues and phrases that the phishing detector identifies as suspicious. While our demonstration suggests that the detectors' assessment matches that of users, this might not be the case for any phishing detector, restricting the applicability of our approach. Second, although we utilized a real-world dataset, a real phishing detector, and included human labelers, our evaluation is nevertheless artificial. With the technical design risks out of the way, an evaluation based on established concepts for the evaluation of user-focused anti-phishing measures is an essential next step. Third, while the design of the explanations was informed by research in phishing susceptibility, our approach in itself does not constitute a full user-focused anti-phishing measure. Further development towards its real-world application will, amongst others, require extensive user interface design. These limitations notwithstanding, our approach provides a first glimpse of the exciting potential of XAI methods for applications in IT security and beyond.

6 Acknowledegments

We kindly thank Rakesh M. Verma (University of Houston) for providing us the dataset.

References

1. O'Donnell, L.: Coronavirus 'Financial Relief' Phishing Attacks Spike, 2020, https://threatpost.com/coronavirus-financial-relief-phishing-spike/154358/ (Accessed: 28.08.2020)
2. Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L.F., Hong, J.: Teaching Johnny Not to Fall for Phish. ACM Trans. Internet Technol. 10 (2010)
3. Parsons, K., Butavicius, M., Pattinson, M., McCormac, A., Calic, D., Jerram, C.: Do Users Focus on the Correct Cues to Differentiate Between Phishing and Genuine Emails? In: 26th Australasían Conference on Information Systems, Adelaide, Australia (2016)
4. Gupta, B.B., Arachchilage, N.A.G., Psannis, K.E.: Defending against phishing attacks: taxonomy of methods, current issues and future directions. Telecommun. Syst. 67, 247–267 (2018)
5. Pienta, D., Thatcher, J., Johnston, A.: A Taxonomy of Phishing: Attack Types Spanning Economic, Temporal, Breadth, and Target Boundaries. In: Proceedings of the 13th Pre-ICIS Workshop on Information Security and Privacy, AIS, San Francisco, CA, USA (2018)
6. Hong, J.: The State of Phishing Attacks. Commun. ACM 55, 74–81 (2012)

7. Khonji, M., Iraqi, Y., Jones, A.: Phishing Detection: A Literature Survey. IEEE Commun. Surv. Tutorials 15, 2091–2121 (2013)
8. Nguyen, C.: Learning Not To Take the Bait: An Examination of Training Methods and Overlerarning on Phishing Susceptibility. PhD thesis. University of Oklahoma, Norman, OK, USA (2018)
9. Albakry, S., Vaniea, K.: Automatic phishing detection versus user training, Is there a middle ground using XAI? In: CEUR Workshop Proceedings, vol. 2151 (2018)
10. Williams, E.J., Hinds, J., Joinson, A.N.: Exploring susceptibility to phishing in the workplace. Int. J. Hum. Comput. Stud. 120, 1–13 (2018)
11. Harrison, B., Svetieva, E., Vishwanath, A.: Individual processing of phishing emails: How attention and elaboration protect against phishing. Online Inf. Rev. 40, 265–281 (2016)
12. Dennis, A.R., Minas, R.K.: Security on Autopilot: Why Current Security Theories Hijack our Thinking and Lead Us Astray. DATABASE Adv. Inf. Syst. 49, 15–38 (2018)
13. Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., Jerram, C.: Phishing for the Truth: A Scenario-Based Experiment of Users' Behavioural Response to Emails. In: SEC 2013, IFIP AICT vol. 405, pp. 366–378, Springer (2013)
14. Blythe, M., Petrie, H., Clark, J.A.: F for Fake: Four Studies on How We Fall for Phish. In: CHI 2011, pp. 3469–3478, ACM, Vancouver, BC, Canada (2011)
15. Vishwanath, A., Herath, T., Chen, R., Wang, J., Rao, H.R.: Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. Decis. Support Syst. 51, 576–586 (2011)
16. Vishwanath, A., Harrison, B., Ng, Y.J.: Suspicion, Cognition, and Automaticity Model of Phishing Susceptibility. Communic. Res. 45, 1146–1166 (2018)
17. Gunning, D.: Explainable Artificial Intelligence (XAI), 2017, https://www.darpa.mil/program/explainable-artificial-intelligence (Accessed: 20.08.2020)
18. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. 51, 1–42 (2019)
19. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Thirty-Second AAAI Conference on Artificial Intelligence, pp. 1527–1535, AAAI, New Orleans, LA, USA (2018)
20. Martens, D., Provost, F.: Explaining Data-Driven Document Classifications. MIS Q. 38, 73–99 (2014)
21. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design Science in Information Systems Research. MIS Q. 28, 75–105 (2004)
22. Jakobsson, M.: The Human Factor in Phishing. In: Priv. Secur. Consum. Inf. (2007)
23. Kim, D., Hyun Kim, J.: Understanding persuasive elements in phishing e-mails. Online Inf. Rev. 37, 835–850 (2013)
24. Zeng, V., Baki, S., Aassal, A. El, Verma, R., Felipe, L., De Moraes, T., Das, A.: Diverse Datasets and a Customizable Benchmarking Framework for Phishing. In: IWSPA '20, pp. 35–41, ACM, New Orleans, LA, USA (2020)
25. Sheng, S., Wardman, B., Warner, G., Cranor, L.F., Hong, J., Zhang, C.: An Empirical Analysis of Phishing Blacklists Steve. In: Sixth Conference on Email Anti-Spam, Mountain View, CA, USA (2009)

26. Verma, R.M., Zeng, V., Faridi, H.: Data Quality for Security Challenges: Case Studies of Phishing, Malware and Intrusion Detection Datasets. In: CCS '19, pp. 2605–2607, ACM, London, UK (2019)
27. Karumbaiah, S., Wright, R.T., Durcikova, A., Jensen, M.L.: Phishing Training: A Preliminary Look at the Effects of Different Types of Training. In: Proceedings of the 11th Pre-ICIS Workshop on Information Security and Privacy, AIS, Dublin, Ireland (2016)
28. Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L.F., Hong, J., Nunge, E.: Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish. In: SOUPS 2007, pp. 88–99, Pittsburgh, PA, USA (2007)
29. Canova, G., Volkamer, M., Bergmann, C., Borza, R.: NoPhish: An Anti-Phishing Education App. In: STM 2014, LNCS vol. 8743, pp. 188–192, Springer (2014)
30. Moody, G.D., Galletta, D.F., Dunn, B.K.: Which phish get caught? An exploratory study of individuals′ susceptibility to phishing. Eur. J. Inf. Syst. 26, 564–584 (2017)
31. Wang, J., Li, Y., Rao, H.R.: Overconfidence in Phishing Email Detection. J. Assoc. Inf. Syst. 17, 759–783 (2016)
32. Volkamer, M., Renaud, K., Reinheimer, B.: TORPEDO: TOoltip-poweRed Phishing Email DetectiOn. In: SEC 2016, IFIP AICT vol. 471, pp. 161–175, Springer (2016)
33. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing Theory-Driven User-Centric Explainable AI. In: CHI 2019, ACM, Glasgow, UK (2019)
34. Lipton, Z.C.: The Mythos of Model Interpretability. Queue 16, 1–27 (2018)
35. Lei, T., Barzilay, R., Jaakkola, T.: Rationalizing Neural Predictions. In: EMNLP 2016, pp. 107–117, ACL, Stroudsburg, PA, USA (2016)
36. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: KDD 2016, pp. 1135–1144, ACM, San Francisco, CA (2016)
37. Lundberg, S., Lee, S.-I.: A Unified Approach to Interpreting Model Predictions. In: NIPS 2017, pp. 4765–4774, Curran Associates, Long Beach, CA, USA (2017)
38. Weerts, H.J.P., van Ipenburg, W., Pechenizkiy, M.: A Human-Grounded Evaluation of SHAP for Alert Processing. In: Proceedings of the KDD Workshop on Explainable AI, Anchorage, AK (2019)
39. Fernandez, C., Provost, F., Han, X.: Counterfactual Explanations for Data-Driven Decisions. In: ICIS 2019, AIS, Munich, Germany (2019)
40. Förster, M., Klier, M., Kluge, K., Sigler, I.: Evaluating Explainable Artificial Intelligence – What Users Really Appreciate. In: ECIS 2020, AIS (2020)
41. Burdisso, S.G., Errecalde, M., Montes-y-Gómez, M.: t-SS3: a text classifier with dynamic n-grams for early risk detection over text streams. arXiv:1911.06147 (2019)
42. Gedikli, F., Jannach, D., Ge, M.: How should I explain? A comparison of different explanation types for recommender systems. Int. J. Hum. Comput. Stud. 72, 367–382 (2014)
43. Ribera, M., Lapedriza, A.: Can we do better explanations? A proposal of User-Centered Explainable AI. In: Joint Proceedings of the ACM IUI 2019 Workshop, ACM, Los Angeles, CA (2019)
44. Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M.F., Eckersley, P.: Explainable machine learning in deployment. In: FAT*20, pp. 648–657, ACM, Barcelona, Spain (2020)

45. Russel, S. and P. Norvig: Artificial Intelligence. 4th edition, Pearson (2020)
46. Kaufmann, E., Kalyanakrishnan, S.: Information complexity in bandit subset selection. J. Mach. Learn. Res. 30, 228–251 (2013)
47. Venable, J., Pries-Heje, J., Baskerville, R.: FEDS: a Framework for Evaluation in Design Science Research. Eur. J. Inf. Syst. 25, 77–89 (2016)
48. Doshi-Velez, F., Kim, B.: Considerations for Evaluation and Generalization in Interpretable Machine Learning. In: Explainable and Interpretable Models in Computer Vision and Machine Learning, pp. 3–17, Springer, Cham (2018)