Beyond the Number: Improving Phishing Risk Perception with Transparent Suspicion Scores

Peifeng He[*†], Kejia Liu[*], Ethan Huang[*], Pardis Emami-Naeini
Duke University

Abstract

E-mail is an important communication tool in our daily lives. However, there are many phishing e-mails to trick recipients into revealing sensitive information like passwords or credit card details, or into clicking malicious links. There are many technical measures to prevent users from trusting the content from phishing e-mails. Among many phishing email warning methods, suspicion score is generally considered to be one of the most effective. However, without knowing the details and explanations behind the scores, a single score may represent different levels of danger in the eyes of different people. We solved this problem by providing a transparent and explanatory representation of the score, which can improve the accuracy and reduce the variance of users' identification of phishing emails.

1 Introduction

E-mail is one of the most popular tools to employ in model life due to the convenience of Internet [16]. Users can send various types information to others easily. Thus, it provides an opportunity for adversaries to trick recipients to transfer their money or expose their sensitive information, which has caused major security issues, psychological damage and financial losses [8, 14, 24, 26, 30]. In addition to that, these phishing emails are often difficult to distinguish [23, 34, 35]

To reduce the negative impact of phishing emails [25], there are many digital users' nudges to prevent users from falling into the trap of adversaries such as: suspicion score [17, 38, 42], passive indicators [17], in-text analysis [22], and active indicators [2, 7]. Suspicion score is a helpful users' nudge technique since it provides a rating of suspiciousness of an email to raise users' awareness of an e-mail [42].

Suspicion score is helpful but it has an obvious shortcoming that it only presents a score [17,42]. The suspicion score is calculated using well-defined methods [21, 34]; however, presenting only a single number makes it appear as a black-box output to the user. Users do not understand the process of its calculation. Thus, when users observe a suspicion score, they do not understand which part of the email is suspicious [34]. This will result in them not being able to accurately identify the risk of the email. Secondly, since there is only one number, different people may have different opinions on the risk level represented by the number, which will lead to a large variance in the judgment made based on this number [27, 40]. Therefore, we propose adding transparent explanation to the common suspicion score. Let users know which factors lead to the high or low score. We believe that a transparent and explanatory representation of the suspicion score can enhance individual risk perception accuracy and reduce variance in phishing email detection among participants. Prior work has shown that explanatory cues are effective in related warning dialogs for suspicious online content [4]. Therefore, this study investigates whether providing transparent explanations alongside suspicion scores enhances users' accuracy and reduces variance in identifying phishing emails.

Usually, suspicion score is calculated based on sender mismatch, request credentials, the sign of urgency, sign of offer, suspicious subject and link mismatch [34]. In order to test whether the improved suspicion score can effectively improve the accuracy of users' judgment of phishing emails, we designed a survey and asked 31 participants to answer a series of questions. In the first part of the questionnaire, they will see five emails and their suspicion scores. Based on the suspicion scores and the email content, they will rate the risk factors of these five emails (ranging from 1 to 10, with higher numbers representing higher risks). At the same time, they will also provide the basis for their judgment on whether each email is risky. In the second part of the questionnaire, we added an analysis of the score on the basis of providing suspicion scores. Participants will score the suspiciousness of the email again and answer whether such an analysis will make their judgment easier. We also collected participants' suspicion scores above what level they would consider the email to be

[*] Equal contribution. [†] Corresponding author: ph162@duke.edu.
This is the paper for PETS 2025 poster "Improving Phishing Risk Perception with Transparent Suspicion Scores".

risky.

We found that with only a suspicion score, users had limited ability to identify the risk factor of an email. Based on the reasoning they provided, participants were not able to accurately identify why the email was risky. In this case, users may give very different risk scores to the same email. After providing an analysis of the score, participants' judgment of the risk factor of the email is closer to the actual situation. And the variance of the scores given by the participant group for the risk factor of the email is smaller. Therefore, we can conclude that after provide a transparent and explanatory suspicion score, the participants' ability to judge the risk factor of the email has been improved.

The refined suspicion score providing detailed factors can bring positive impact on email service providers and security researchers, each of whom plays a unique role in mitigating the risks of phishing and improving user security. Suspicion score equipped with detailed factors' calculation enables email service providers to design more features based on these factors to alert users. Also, email service providers can implement their detection systems to stop the spread of phishing emails at the first place. Meanwhile, security researchers can learn more about the use of detailed explanation of suspicion score, which can be a trigger for other security measures.

2 Background

2.1 User-centric nudge

User-centric nudges are interventions aimed at guiding user behavior toward safer decisions by subtly adjusting the context or environment in which choices are made [5, 32, 36]. Such nudges, including warnings and visual cues [1, 19, 31], can significantly improve security outcomes. Prior research has shown various ways of effectively nudging users towards safer cybersecurity practices. For instance, training and advice that utilize stories, relatable scenarios, or peer comparisons have proven particularly effective [13, 39]. These approaches align with broader human-centered security design frameworks that emphasize the importance of cognitive load, risk communication, and behavioral nudging [9, 12].

2.2 Suspicion score

This score quantifies the likelihood that a given email is malicious, enabling automated systems and users to assess and respond to potential threats effectively. The development and refinement of suspicion scores have been instrumental in enhancing the accuracy and efficiency of phishing detection mechanisms. Previous literature highlights the critical role of transparent and interpretable suspicion scoring methods to enhance user trust and usability in phishing detection [15]. Recent research reinforces this point, demonstrating that real-time, evidence-based alerts significantly enhance users' ability to detect phishing by combining high-performing classifiers with interpretable cues [6]. Similarly, feature engineering methods and gaze-based cognitive studies have validated the importance of explainability and user-centric presentation of suspicious features in emails [18, 28, 29].

2.3 AI email analysis models

Since ChatGPT can provide detailed numerical outputs for individual factors through sentiment analysis [41], we leveraged this capability to compute and present the percentage contribution of each factor associated with a given phishing email. By explicitly quantifying these contributing elements in a structured and interpretable format, we aim to reduce user uncertainty, mitigate bias in interpretation, and ultimately enhance the transparency and consistency of the suspicion score evaluation process.

3 Methodology

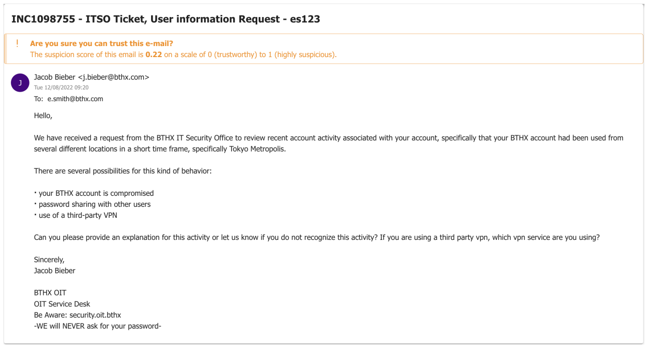
We hope to understand whether participants can identify risky emails with the help of the original and improved suspicion scores, and whether they can determine where the risk comes from. We use both qualitative and quantitative results to determine whether transparent and explanatory suspicion scores can positively affect participants' risk perception of emails. In this section, we will introduce how our study was set up, the design details of our study, and the feedback and analysis of participants' answers.

3.1 Participants

We recruited participants using a targeted social media campaign on platforms such as Instagram and WeChat, where we shared recruitment messages in relevant online communities. The posts included a brief description of the study, eligibility criteria, and a link to a Qualtrics-based screening form. Participants were required to meet all inclusion criteria to proceed, which included being at least 18 years old and regularly using email for personal or professional communication. Individuals with professional experience or an academic background in cybersecurity were excluded to avoid bias. Participants who completed the full study were entered into a randomized drawing to win $30, with eligibility contingent on answering all survey questions; no partial compensation was offered.

3.2 Tasks

We conducted an online user study via a Qualtrics survey to evaluate the usability of phishing warnings and the effectiveness of user nudging mechanisms through numerical suspicion scores and contextual explanations. The main objective was to understand how individuals perceive phishing-related risks when presented with automated suspicion indicators,



INC1098755 - ITSO Ticket, User information Request - es123

! **Are you sure you can trust this e-mail?**
The suspicion score of this email is **0.22** on a scale of 0 (trustworthy) to 1 (highly suspicious).

Jacob Bieber <j.bieber@bthx.com>
Tue 12/08/2022 09:20
To: e.smith@bthx.com

Hello,

We have received a request from the BTHX IT Security Office to review recent account activity associated with your account, specifically that your BTHX account had been used from several different locations in a short time frame, specifically Tokyo Metropolis.

There are several possibilities for this kind of behavior:

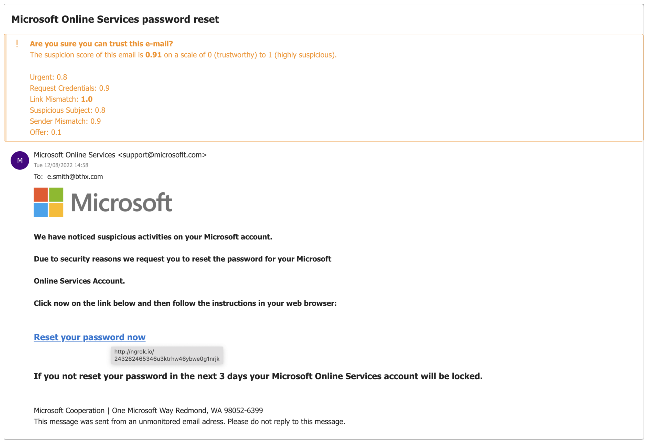
- your BTHX account is compromised
- password sharing with other users
- use of a third-party VPN

Can you please provide an explanation for this activity or let us know if you do not recognize this activity? If you are using a third party vpn, which vpn service are you using?

Sincerely,
Jacob Bieber

BTHX OIT
OIT Service Desk
Be Aware: security.oit.bthx
-WE will NEVER ask for your password-

Figure 1: an email with suspicion score only



Microsoft Online Services password reset

! **Are you sure you can trust this e-mail?**
The suspicion score of this email is **0.91** on a scale of 0 (trustworthy) to 1 (highly suspicious).

Urgent: 0.8
Request Credentials: 0.9
Link Mismatch: **1.0**
Suspicious Subject: 0.8
Sender Mismatch: 0.9
Offer: 0.1

Microsoft Online Services <support@microsoflt.com>
Tue 12/08/2022 14:58
To: e.smith@bthx.com

We have noticed suspicious activities on your Microsoft account.

Due to security reasons we request you to reset the password for your Microsoft

Online Services Account.

Click now on the link below and then follow the instructions in your web browser:

Reset your password now

http://ngrok.io/
243262465346u3ktrhw46ybwe0g1nrjk

If you not reset your password in the next 3 days your Microsoft Online Services account will be locked.

Microsoft Cooperation | One Microsoft Way Redmond, WA 98052-6399
This message was sent from an unmonitored email adress. Please do not reply to this message.

Figure 2: an email with suspicion score and transparent explanations

suspicion score, and how explanatory feedback might nudge users toward more accurate threat assessments.

In the survey, we will provide five emails to participants. In order to simulate real life or work scenarios, we ask participants to handle these five emails as E. Smith, an employee of a virtual company called "BTHX".

In order to simulate the real-world situation as much as possible, our five emails have different characteristics. The first one is from the technical department of BTHX, prompting E. Smith's associated account to have activities in different locations around the world. The suspicion score of this email is 0.22. The second email comes from an external website to remind E. Smith to claim the prize, which contains an external link that requires E. Smith to provide additional information. The suspicion score of this email is 0.9. The third email comes from a disguised and very hidden Microsoft account, prompting E. Smith's Microsoft account activity is abnormal and the password needs to be reset. The suspicion score of this email is 0.91. The fourth email comes from a disguised UPS. In the email, the other party provides the correct tracking number, but provides forged contact information. The suspicion score of this email is 0.6. The fifth email is forwarded from a company employee, which includes a '.pdf.exe' file that needs to be viewed by E. Smith. The suspicion score of this email is 0.75. Please refer to Table 1 for specific email information. The standard of emails we collected is that we hope that each email contains topics that are frequently encountered in life or work. One of the emails is a normal email (the first one). In the other four emails, we incorporated many disguises to increase the difficulty for participants to identify the risky elements.

Our emails were presented to participants using a mock inbox that simulated the Outlook UI, since previous research validates the effectiveness of realistic interfaces to elicit authentic user responses to phishing attempts [10, 42].

The study followed a two-phase experimental design. In the first phase, participants were shown five email screenshots (figure 1, see rest of the screenshots in the appendix), each accompanied by a minimalist phishing warning in the form of a numerical suspicion score (from 0 to 1). Participants then rated the likelihood of each email being a phishing attempt on a 10-point Likert scale, simulating real-world decision-making under uncertainty. After rating each email, participants were asked to provide the factors that led to their rating. These factors included: suspicious sender, unexpected request for information, poor grammar/spelling, urgency or threatening language, and email formatting issues. We wanted to see if participants could accurately identify issues in an email given only a suspicion score. After this part, participants had some understanding of what a suspicion score is and had some initial experience identifying phishing emails. We asked participants to write above what level they would consider the email to be risky.

In the second phase, the same emails were presented again, now with transparent explanations of the suspicion score (figure 2). These explanations highlighted six contributing phishing factors: sender mismatch, credential requests, urgency, offers, suspicious subject, and link mismatches [34]. Each paired with a GPT-generated qualitative score indicating its severity or presence and the overall suspicion score was derived from the combined intensity of these factors [33]. This breakdown served as a user-centric nudge, helping participants better understand the rationale behind the phishing assessment and make more informed judgments. After assigning their scores, we asked participants to indicate whether the transparent explanations helped them make their judgments on the risk level of the emails.

3.3 Data Processing

We processed the Likert-scale responses to quantify participant judgments on phishing likelihood across both phases of the study. For each question (Q1–Q5 for the initial phase without explanations and Q6–Q10 for the phase with breakdowns),

Table 1: Summary of the five e-mails included in the survey

E-mail index	Sender	Major phishing characteristics	Summery	Suspicion score
1	j.bieber@bthx.com	Not a phishing e-mail	Request for user to explain unusual account activity, possibly due to compromise, sharing, or VPN usage.	0.22
2	m.greggs@tucsonweekend.com	Suspicious subject, offer	Claim your free gift within 24 hours by verifying your account through the provided link.	0.9
3	support@microsoflt.com	Link mismatch, request credentials, urgent	User is urged to reset Microsoft password due to suspicious activity or risk account being locked.	0.91
4	uscustsvc@ups.com	Sender mismatch	User is informed their UPS package couldn't be shipped due to registration error; account verification requested.	0.6
5	m.locke@bthx.com	Link mismatch	User is informed their UPS package couldn't be shipped due to registration error; account verification requested.	0.75

we first extracted the numerical score from the textual Likert responses using a regular expression parser.

To visualize the spread and identify outliers, we plotted boxplots for each question. The distributions revealed several extreme values, so we applied outlier filtering using Z-scores: only responses with $|z| < 3$ were retained. This step improved the reliability of our aggregated statistics.

3.4 Data Analysis

For the data from the first section only, we first determine how many users gave the correct answer based on the answers they provided when there was only one suspicion score. In our question, for risky emails (emails 2, 3, 4, and 5), if the user provided a value greater than neutral, it was considered to have correctly identified the risk of the email. For emails that were not dangerous (email 1), we considered a correct score of 1 to 5 to be the correct range of identification. On this basis, since we also collected the reasons why the emails were risky provided by users, we further counted how many participants in the group who correctly identified the risk factor of the email also provided the correct basis. And for the level they would consider the email to be risky, We show the distribution based on the data.

After we cleaned the data for both sections, we computed the mean and standard deviation for each question to assess shifts in participant perception. To evaluate the effect of explanation breakdowns, we paired and compared each email's responses before and after the inclusion of explanations—specifically, Q1 with Q6, Q2 with Q7, Q3 with Q8, Q4 with Q9, and Q5 with Q10. We aimed to determine whether the mean response in the second phase moved closer to the GPT-generated suspicion score, which we treated as a proxy for ground truth. Additionally, we examined whether the standard deviation decreased, indicating stronger participant consensus and suggesting that the breakdowns helped reduce ambiguity in interpreting the phishing risk. Such comparative analytical methods have been demonstrated as effective and robust for evaluating cybersecurity interventions [20]. To support our analysis, we generated histograms for the computed means and standard deviations across all questions, enabling a visual comparison of central tendencies and variability before and after introducing the breakdowns.

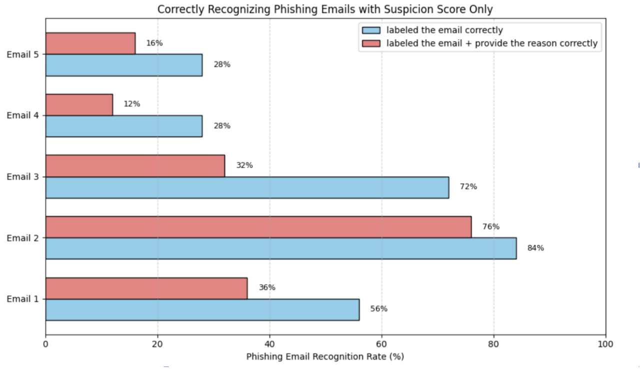


Figure 3: The percentage of emails and phishing reasons correctly identified

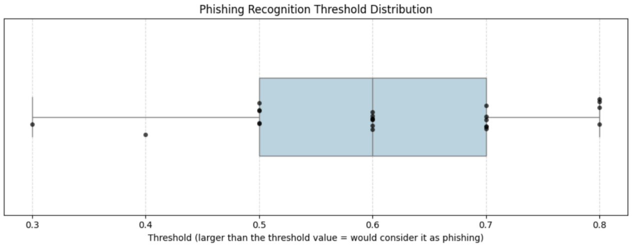


Figure 4: Distribution of risk threshold

4 Results

4.1 With suspicion score only

We found that when there is only one score, participants have limited ability to identify whether an email is risky. When the risk of the email is hard to identify (email 4 and 5), the recognition ability of participants is significantly reduced. And only a smaller number of people can provide correct evidence even if they labeled the email correctly (Figure 3). In everyone's threshold distribution diagram (Figure 4), we found that everyone's answers are very scattered. We can see that a person think that 0.3 or higher means it is risky, while some other participants think that the email is risky only if the suspicion score is 0.8 or higher. From the first part of the data collected with suspicion score only, we found that the ability of participants to correctly identify the risk of the email is very limited. And because different people have different understandings of the score, participants have very different views on the potential dangers represented by the same score.

4.2 Improved accuracy

We mentioned in the previous section that we calculated the average of the participants' scores for each question. We

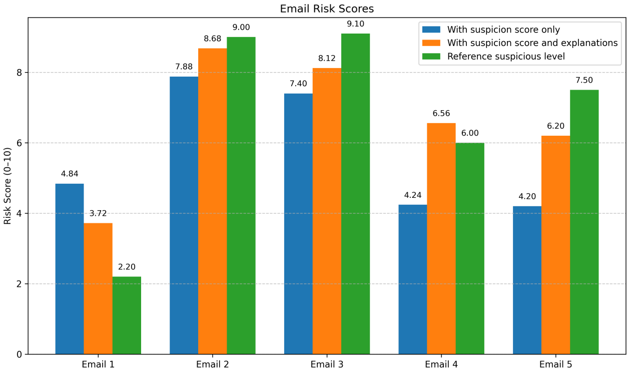


Figure 5: Average of scores before and after detailed explanations are added for 5 emails, comparing with the ground truth score

hope to determine whether the participants have a clearer understanding of the risk of the email by the change in the average of the corresponding question. Since our suspicion score is calculated from 0 to 1, and the risk factor of the email given by the participants is from 1 to 10, in order to unify the standards, the risk level of the email is equal to the suspicion score of the email times 10. Since the suspicion scores for the five emails are $2.2, 0.9, 0.91, 0.6, 0.75$ respectively, the risk level of the five emails that will be used in the comparison will be $2.2, 9, 9.1, 6, 7.5$ respectively.

As displayed in Figure 6, for each email, when there is a transparent explanation, the participants' perception of the email risk factor is closer to the actual risk factor of the email. Here, closer means that for non-phishing emails, the average score has decreased, and people are more confident in the security of this email. For emails with high risk levels, the increase in the average score means that people have a clearer understanding of the risks of these emails. In both cases, users are more capable of identifying the risk factor of emails. Therefore, we believe that suspicion score with transparent explanation can improve the accuracy of public identification of phishing emails.

4.3 Decreased variance

We also calculated the standard deviation of the scores given by participants. Our expectation is that with transparent explanations, people are more likely to agree on the risk factor of an email. In other words, we hope that with the presentation of a more detailed suspicion score calculation method, the standard deviation of users' assessment of the risk factor of an email should decrease.

To test our hypothesis, we calculated the STD of the risk scores of those five emails before and after, as shown in figure 6). Before providing transparent explanations, the STDs

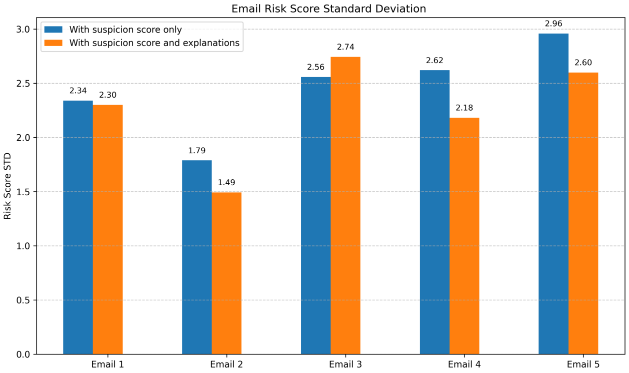


Figure 6: Standard deviation of scores before and after detailed explanations are added for 5 emails

of the five emails were 2.34, 1.787, 2.566, 2.619, and 2.958. After providing transparent explanations, the STDs of the five emails were 2.30, 1.49, 2.73, 2.18, and 2.98. Except for the third email where the STD increased, the STDs of the other four emails decreased. We believe that the reason for the increase in STD in the third email may be that some participants did not notice the links introduced by the curser. For the two emails that did not have very high suspicion scores but were phishing emails, the standard deviation decreased most significantly. In general, due to the decrease in STD, participants' risk perception of emails became more unified and they were more inclined to reach a consensus.

5 Discussion

Comparing with earlier studies, we observed a key difference: providing explanatory information further enhanced users' judgment accuracy and reduced individual variation. This contrasts with earlier studies that treated suspicion scores as standalone indicators.

We attribute this improvement to the added transparency, which helps users better understand the rationale behind the score, thereby building trust and supporting more informed decisions. This suggests that explanation-aware design can significantly improve the usability and effectiveness of phishing detection systems.

5.1 Limitations

First, while our sample of 31 participants provided meaningful insights, it remains relatively small and may not represent the broader population. Future studies should consider a larger and more diverse participant pool, including users of different age groups, technical backgrounds, and email usage patterns.

Second, participants reviewed emails in a simulated environment rather than a live email platform, which may not perfectly reflect real-world behavior [11]. For instance, in an actual email client, users might interact differently with elements like links, attachments, or sender metadata. This could especially affect nuanced interpretations such as the one observed in Email 3, where the standard deviation unexpectedly increased after explanations were added.

Third, our suspicion score explanations were static and did not allow participants to interact with or explore the underlying indicators. Adding interactivity—such as clicking to highlight risky phrases or preview links—might further improve user understanding and engagement [37].

Last but not least, a large portion of participants have prior knowledge of phishing and had received some form of education about cybersecurity. As a result, their responses may not be fully representative of the general population. This introduces a degree of sampling bias, limiting the generalization of our findings to broader user groups, particularly those with little to no cyber-security awareness.

5.2 Future Work

There are several directions for expanding this research:

- **Interactive Explanations:** Investigate the effectiveness of interactive explanation interfaces where users can explore how the suspicion score is computed in real-time, e.g., hovering over suspicious phrases or link mismatches.

- **Personalized Nudges:** Explore whether adapting explanations based on user behavior or expertise (e.g., tech-savvy vs. non-technical users) leads to better outcomes.

- **Longitudinal Studies:** Conduct follow-up studies to examine whether the understanding gained through transparent scores leads to long-term improvement in phishing detection accuracy.

- **Integration with Email Clients:** Test the practical usability and performance of suspicion score explanations when deployed directly inside real email platforms like Outlook or Gmail, ideally through a browser extension or plugin.

- **Automated Feedback Loops:** Examine whether users can provide feedback on the explanations themselves to help models learn which types of explanations are most useful or trustworthy.

Ultimately, combining accurate scoring models with transparent and human-centered explanations can build user trust and mitigate risks more effectively than opaque systems alone.

6 Conclusion

Phishing remains one of the most prevalent cyber threats, and while automated systems can detect many attacks, the end user still plays a crucial role in stopping targeted attempts [3]. In this work, we showed that supplementing suspicion scores with transparent, factor-based explanations improves user performance in identifying phishing emails. Participants not only rated emails more accurately, but also showed more consensus in their risk assessments, suggesting that explanatory nudges help reduce ambiguity.

These findings suggest that email security systems should consider going beyond raw scores and invest in interpretable outputs that guide user judgment. As phishing techniques grow more sophisticated, empowering users with the right information—not just numbers—will be key to maintaining secure communication.

References

[1] Devdatta Akhawe and Adrienne Porter Felt. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *22nd USENIX Security Symposium (USENIX Security '13)*, USENIX Security '13, pages 257–272, Washington, D.C., 2013. USENIX Association.

[2] Rana Alabdan. Phishing attacks survey: Types, vectors, and technical approaches. *Future Internet*, 12(10), 2020.

[3] Zainab Alkhalil, Chaminda Hewage, Liqaa Nawaf, and Imtiaz Khan. Phishing attacks: A recent comprehensive study and a new anatomy. *Frontiers in Computer Science*, 3, 2021.

[4] Areej Alnajim and Malcolm Munro. Explanations in warning dialogs to help users defend against phishing attacks. In *International Journal of Human-Computer Studies*, volume 170, page 102977. Elsevier, 2023.

[5] Aurélien Baillon, Jeroen De Bruin, Aysil Emirmahmutoglu, Evelien Van De Veer, and Bram Van Dijk. Informing, simulating experience, or both: A field experiment on phishing risks. In *PLoS ONE*, volume 14 of *PLoS ONE*, USA, 2019. Public Library of Science.

[6] Shahryar Baki, Fatima Zahra Qachfar, Rakesh M. Verma, Ryan Kennedy, and Daniel N. Jones. Real-time, evidence-based alerts for protection from phishing attacks. *IEEE Transactions on Dependable and Secure Computing*, 22(2):1055–1069, 2025.

[7] Dennik Baltuttis and Timm Teubner. Effects of visual risk indicators on phishing detection behavior: An eye-tracking experiment. *Computers & Security*, 144:103940, 2024.

[8] Marzieh Bitaab, Haehyun Cho, Adam Oest, Penghui Zhang, Zhibo Sun, Rana Pourmohamad, Doowon Kim, Tiffany Bao, Ruoyu Wang, Yan Shoshitaishvili, et al. Scam pandemic: How attackers exploit public fear through phishing. In *APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–10, 2020. IEEE.

[9] Jim Blythe, Jean Camp, and Vaibhav Garg. Targeted risk communication for computer security. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, IUI '11, page 295–298, New York, NY, USA, 2011. Association for Computing Machinery.

[10] Cristian Bravo-Lillo, Saranga Komanduri, Lorrie Faith Cranor, Robert W. Reeder, Manya Sleeper, Julie Downs, and Stuart Schechter. Your attention please: designing security-decision uis to make genuine risks harder to ignore. SOUPS '13, New York, NY, USA, 2013. Association for Computing Machinery.

[11] Amy G. Buhler, Brittany Brannon, Tara Tobin Cataldo, Ixchel M. Faniel, Lynn Silipigni Connaway, Joyce Kasman Valenza, Rachael Elrod, and Christopher Cyr. How real is real enough? participant feedback on a behavioral simulation used for information-seeking behavior research. *Journal of Librarianship and Information Science*, 55(1):191–207, 2023.

[12] Casey Inez Canfield, Baruch Fischhoff, and Alex Davis. Quantifying phishing susceptibility for detection and behavior decisions. *Human Factors*, 58(8):1158–1172, 2016. PMID: 27562565.

[13] Deanna D. Caputo, Shari Lawrence Pfleeger, Jesse D. Freeman, and M. Eric Johnson. Going spear phishing: Exploring embedded training and awareness. *IEEE Security Privacy*, 12(1):28–38, 2014.

[14] Xi Chen, Indranil Bose, Alvin Chung Man Leung, and Chenhui Guo. Assessing the severity of phishing attacks: A hybrid data mining approach. *Decision Support Systems*, 50(4):662–672, 2011.

[15] Lorrie Faith Cranor. A framework for reasoning about the human in the loop. UPSEC'08, USA, 2008. USENIX Association.

[16] Laura A Dabbish, Robert E Kraut, and Susan R Fussell. Understanding email use: predicting action on a message. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 691–700, 2005.

[17] Giuseppe Desolda, Francesco Di Nocera, Lauren Ferro, Rosa Lanzilotti, Piero Maggi, and Andrea Marrella. *Alerting Users About Phishing Attacks*, pages 134–148. 06 2019.

[18] Andrei Dumitras, Cristinel Mihai Mocan, and Ciprian Oprisa. A feature engineering approach for detecting phishing emails. In *2024 IEEE 20th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 1–8, 2024.

[19] Serge Egelman, Lorrie Cranor, and Jason Hong. You've been warned: An empirical study of the effectiveness of web browser phishing warnings. In *SIGCHI Conference on Human Factors in Computing Systems (CHI)*, CHI '08. ACM, 2008.

[20] Serge Egelman and Eyal Peer. Scaling the security wall: Developing a security behavior intentions scale (sebis). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 2873–2882, New York, NY, USA, 2015. Association for Computing Machinery.

[21] Chibuike Samuel Eze and Lior Shamir. Analysis and prevention of ai-based phishing email attacks. *Electronics*, 13(10), 2024.

[22] Luisa Franchina, Serena Ferracci, and Federico Palmaro. Detecting phishing e-mails using text mining and features analysis. 12 2021.

[23] Danilo Gentile, Aniello Castiglione, Roberto De Prisco, and Francesco Palmieri. The human factor in phishing: Collecting and analyzing user-centric data for the design of effective mitigations. *Computers Security*, 130:102949, 2023.

[24] Jason Hong. The state of phishing attacks. *Commun. ACM*, 55(1):74–81, January 2012.

[25] Tom N Jagatic, Nathaniel A Johnson, Markus Jakobsson, and Filippo Menczer. Social phishing. *Communications of the ACM*, 50(10):94–100, 2007.

[26] Jurjen Jansen and Rutger Leukfeldt. Coping with cybercrime victimization: An exploratory study into impact and change. *Journal of Qualitative Criminal Justice and Criminology*, 6(2):205–228, 2018.

[27] Hye-Jin Paek and Thomas Hove. Risk perceptions and risk characteristics, 03 2017.

[28] Francesco Pietrantonio, Alessio Botta, Stefania Zinno, Giorgio Ventre, Luigi Gallo, Laura Mancuso, and Roberta Presta. A gaze-based analysis of human detection of email phishing. In *2024 Silicon Valley Cybersecurity Conference (SVCC)*, pages 1–8, 2024.

[29] Krishnaiahgari Karthik Reddy, G. Jaspher W Kathrine, and Dasari Kishan Kumar. Cyber sentinel: Intelligent phishing url identification system employing machine learning methods. In *2024 8th International Conference on Inventive Systems and Control (ICISC)*, pages 168–173, 2024.

[30] J. Roberts. Exclusive: Facebook and google were victims of $100 m payment scam. *Fortune Magazine*, page 27, 2017.

[31] Stuart E. Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. The emperor's new security indicators. In *2007 IEEE Symposium on Security and Privacy (SP '07)*, IEEE S&P '07, pages 51–65. IEEE, 2007.

[32] Christoph Schneider, Markus Weinmann, and Jan vom Brocke. Digital nudging: guiding online user choices through interface design. *Commun. ACM*, 61(7):67–73, June 2018.

[33] Federico Seijo. Squidshing: Analyzes emails for phishing risks. *OpenAI*, 2025. GPT model, `https://chat.openai.com/g/g-8JrlEnLEj-squidshing`.

[34] Kuldeep Singh, Palvi Aggarwal, Prashanth Rajivan, and Cleotilde Gonzalez. What makes phishing emails hard for humans to detect? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 64(1):431–435, 2020.

[35] Michelle P. Steves, Kristen K. Greene, and Mary F. Theofanos. A phish scale: Rating human phishing message detection difficulty. In *Proceedings of the NDSS Symposium*. Internet Society, 2019.

[36] Richard H. Thaler and Cass R. Sunstein. Nudge: Improving decisions about health, wealth, and happiness. Nudge Series, New Haven, CT, 2008. Yale University Press.

[37] Anthony Vance, Jeffrey L. Jenkins, Bonnie Brinton Anderson, and David Bjornn. Fear appeals and information security behaviors: An empirical study. In *MIS Quarterly*. Management Information Systems Research Center, University of Minnesota, 2023. Forthcoming.

[38] Arun Vishwanath, Brynne Harrison, and Yu Ng. Suspicion, cognition, and automaticity model of phishing susceptibility. *Communication Research*, 45:1146–1166, 12 2018.

[39] Rick Wash and Molly M. Cooper. Who provides phishing training? facts, stories, and people like me. CHI '18, page 1–12, New York, NY, USA, 2018. Association for Computing Machinery.

[40] Sherry Ruan Wu, Joannie Rick, and Robert C Miller. Suspicion: A usable email interface for helping users identify suspicious emails. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM, 2020.

[41] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. Sentiment analysis in the era of large language models: A reality check, 2023.

[42] Sarah Y. Zheng and Ingolf Becker. Checking, nudging or scoring? evaluating e-mail user security tools. In *USENIX Symposium on Usable Privacy and Security (SOUPS)*, pages 57–76, 2023. https://www.usenix.org/conference/soups2023/presentation/zheng.

Appendix

A. Consent Form

Thank you for your interest in our study. We are researchers at Duke University conducting a study to explore how individuals interpret and interact with numerical indicators in decision-making contexts related to email content.

Procedures: You will be asked to complete an online survey, including questions about email usage, risk perception, and phishing detection. It will take approximately 10–15 minutes. Your responses will be anonymous.

Participant Requirements:

- Must be 18 years or older
- Must regularly use email for communication
- Must not have professional experience in cybersecurity

Compensation: Upon successful completion, you will be entered into a $30 lottery.

Confidentiality: No identifiable data is collected. Responses are securely stored on Duke-approved encrypted cloud storage (Duke Box).

Voluntariness: Participation is voluntary. You may withdraw at any time.

Consent: By continuing, you confirm:

1. You are 18 years or older
2. You have read and understood this form
3. You voluntarily agree to participate

B. Screening Question

> Q: Do you use email regularly for personal or professional communication?
> A: **Yes / No**

C. Focus Group Scenario (Main Task Context)

Assume your name is E. Smith, and you are checking your new incoming emails. You work for a company called BTHX, where anyone with an @bthx.com email address is a colleague from the same organization. However, their familiarity with security concepts may vary.

You have an order waiting to be shipped by UPS. The tracking number is 1Z 623 X56 03 1224 5784.

D. Recruitment Text

> **Subject: Participate in a Short Research Survey on Email Security!**
> Are you 18+ and use email regularly? Take our 10–15 min anonymous survey about email risk and phishing detection. You could win $30! No cybersecurity background required.
> Click here to begin: https://duke.qualtrics.com/jfe/form/SV_8e3T2DMcFOo0nHM

E. Main Study Questions

Phase 1: Email with Suspicion Score Only

Email 1 – Suspicion Score Only (See Figure 7)

- Q1: How likely is this email a phishing attempt? [1–10 Likert Scale]
- Q1a: What factors influenced your decision? (Check all that apply)
  - Suspicious sender
  - Unexpected request for information
  - Poor grammar/spelling
  - Urgency or threatening language
  - Email formatting issues
  - Nothing seemed suspicious; the email appeared safe
  - Other (please specify): ____

Email 2 – Suspicion Score Only (See Figure 8)

- Q2: Likelihood of phishing? [1–10]
- Q2a: Influencing factors? (Same options as Q1a)

Email 3 – Suspicion Score Only (See Figure 9)

- Q3 and Q3a as above

Email 4 – Suspicion Score Only (See Figure 10)

- Q4 and Q4a as above

Email 5 – Suspicion Score Only (See Figure 11)

- Q5 and Q5a as above

Suspicion Score Threshold

Q(Threshold): Above what score (0–1) would you consider an email suspicious?

Phase 2: Email with Transparent Explanation

The following emails include detailed factor-based breakdowns for the suspicion score:

- **Sender mismatch:** Triggered when the sender's name or domain appears inconsistent with who they claim to be (e.g., spoofed names, typos).
- **Request credentials:** Present when the email asks for personal or confidential data such as passwords, account info, or verification codes.
- **Urgency:** Signaled by time pressure, fear tactics, or threats (e.g., "act now").
- **Offer:** The email promises a gift, help, or reward in a suspicious way.
- **Suspicious subject:** Subject line contains urgency, threats, or unusual punctuation.
- **Link mismatch:** The visible link and actual destination don't match, or an IP address is used.

As you proceed, you'll see a numerical suspicion score and these factors visually highlighted for each email.

Email 1 – Transparent Explanation (See Figure 12)

- Q6: Likelihood of phishing? [1–10]
- Q6a: Did the transparent explanation help you decide?
  - Yes, it made it clearer
  - Somewhat, but I still relied on my judgment
  - No, it didn't make a difference

Email 2 – Transparent Explanation (See Figure 13)

- Q7 and Q7a as above

Email 3 – Transparent Explanation (See Figure 14)

- Q8 and Q8a as above

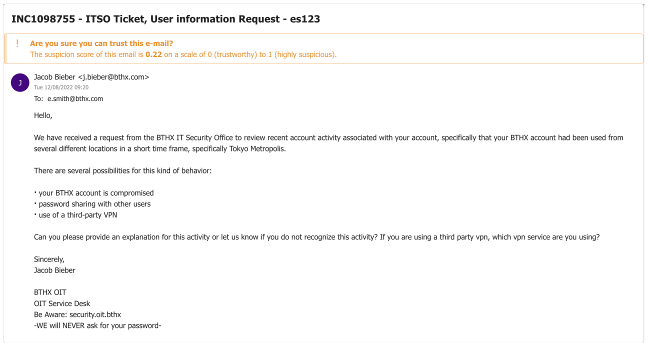
Email 4 – Transparent Explanation (See Figure 15)

- Q9 and Q9a as above

Email 5 – Transparent Explanation (See Figure 16)

- Q10 and Q10a as above

F. Email Screenshots



INC1098755 - ITSO Ticket, User information Request - es123

! **Are you sure you can trust this e-mail?**
The suspicion score of this email is **0.22** on a scale of 0 (trustworthy) to 1 (highly suspicious).

Jacob Bieber <j.bieber@bthx.com>
Tue 12/08/2022 09:20
To: e.smith@bthx.com

Hello,

We have received a request from the BTHX IT Security Office to review recent account activity associated with your account, specifically that your BTHX account had been used from several different locations in a short time frame, specifically Tokyo Metropolis.

There are several possibilities for this kind of behavior:

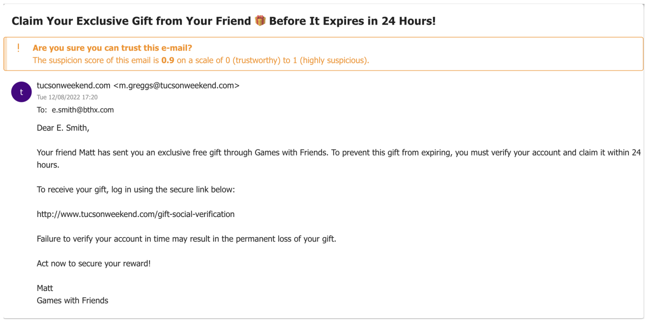
- your BTHX account is compromised
- password sharing with other users
- use of a third-party VPN

Can you please provide an explanation for this activity or let us know if you do not recognize this activity? If you are using a third party vpn, which vpn service are you using?

Sincerely,
Jacob Bieber

BTHX OIT
OIT Service Desk
Be Aware: security.oit.bthx
-WE will NEVER ask for your password-

Figure 7: Email 1 – Suspicion Score Only



Claim Your Exclusive Gift from Your Friend 🎁 Before It Expires in 24 Hours!

! **Are you sure you can trust this e-mail?**
The suspicion score of this email is **0.9** on a scale of 0 (trustworthy) to 1 (highly suspicious).

tucsonweekend.com <m.greggs@tucsonweekend.com>
Tue 12/08/2022 17:20
To: e.smith@bthx.com

Dear E. Smith,

Your friend Matt has sent you an exclusive free gift through Games with Friends. To prevent this gift from expiring, you must verify your account and claim it within 24 hours.

To receive your gift, log in using the secure link below:

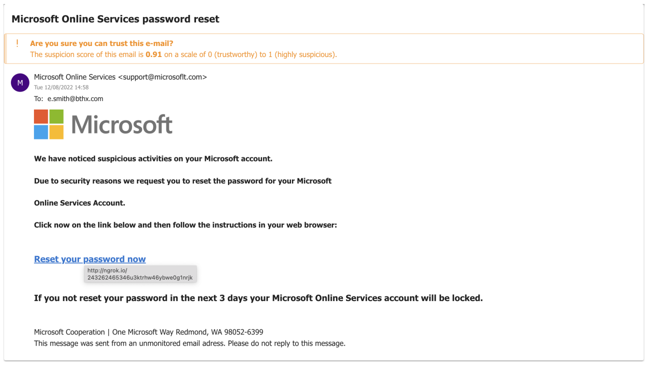
http://www.tucsonweekend.com/gift-social-verification

Failure to verify your account in time may result in the permanent loss of your gift.

Act now to secure your reward!

Matt
Games with Friends

Figure 8: Email 2 – Suspicion Score Only



Microsoft Online Services password reset

! **Are you sure you can trust this e-mail?**
The suspicion score of this email is **0.91** on a scale of 0 (trustworthy) to 1 (highly suspicious).

Microsoft Online Services <support@microsoflt.com>
Tue 12/08/2022 14:58
To: e.smith@bthx.com

Microsoft

We have noticed suspicious activities on your Microsoft account.

Due to security reasons we request you to reset the password for your Microsoft

Online Services Account.

Click now on the link below and then follow the instructions in your web browser:

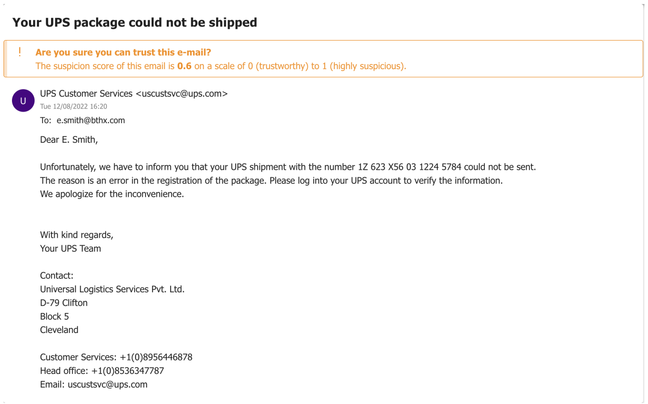
Reset your password now

http://ngrok.io/
243262465346u3ktrhw46ybwe0g1nrjk

If you not reset your password in the next 3 days your Microsoft Online Services account will be locked.

Microsoft Cooperation | One Microsoft Way Redmond, WA 98052-6399
This message was sent from an unmonitored email adress. Please do not reply to this message.

Figure 9: Email 3 – Suspicion Score Only



Your UPS package could not be shipped

! **Are you sure you can trust this e-mail?**
The suspicion score of this email is **0.6** on a scale of 0 (trustworthy) to 1 (highly suspicious).

UPS Customer Services <uscustsvc@ups.com>
Tue 12/08/2022 16:20
To: e.smith@bthx.com

Dear E. Smith,

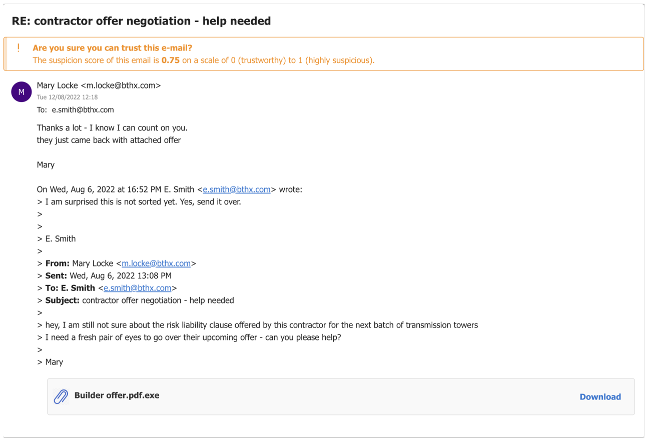
Unfortunately, we have to inform you that your UPS shipment with the number 1Z 623 X56 03 1224 5784 could not be sent.
The reason is an error in the registration of the package. Please log into your UPS account to verify the information.
We apologize for the inconvenience.

With kind regards,
Your UPS Team

Contact:
Universal Logistics Services Pvt. Ltd.
D-79 Clifton
Block 5
Cleveland

Customer Services: +1(0)8956446878
Head office: +1(0)8536347787
Email: uscustsvc@ups.com

Figure 10: Email 4 – Suspicion Score Only



RE: contractor offer negotiation - help needed

! **Are you sure you can trust this e-mail?**
The suspicion score of this email is **0.75** on a scale of 0 (trustworthy) to 1 (highly suspicious).

Mary Locke <m.locke@bthx.com>
Tue 12/08/2022 12:18
To: e.smith@bthx.com

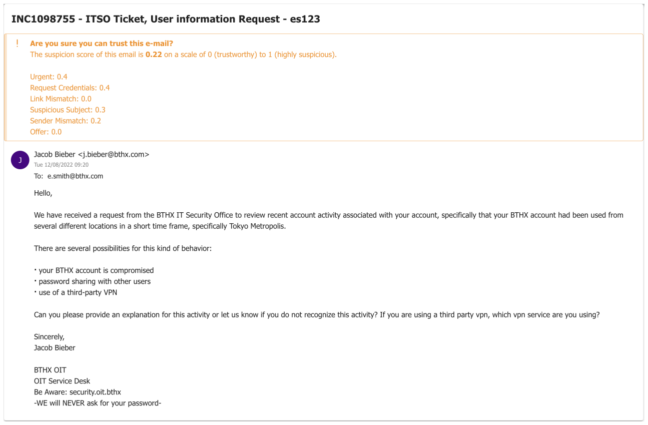
Thanks a lot - I know I can count on you.
they just came back with attached offer

Mary

On Wed, Aug 6, 2022 at 16:52 PM E. Smith <e.smith@bthx.com> wrote:
> I am surprised this is not sorted yet. Yes, send it over.
>
>
> E. Smith
>
> **From:** Mary Locke <m.locke@bthx.com>
> **Sent:** Wed, Aug 6, 2022 13:08 PM
> **To: E. Smith** <e.smith@bthx.com>
> **Subject:** contractor offer negotiation - help needed
>
> hey, I am still not sure about the risk liability clause offered by this contractor for the next batch of transmission towers
> I need a fresh pair of eyes to go over their upcoming offer - can you please help?
>
> Mary

Builder offer.pdf.exe **Download**

Figure 11: Email 5 – Suspicion Score Only



INC1098755 - ITSO Ticket, User information Request - es123

! **Are you sure you can trust this e-mail?**
The suspicion score of this email is **0.22** on a scale of 0 (trustworthy) to 1 (highly suspicious).

Urgent: 0.4
Request Credentials: 0.4
Link Mismatch: 0.0
Suspicious Subject: 0.3
Sender Mismatch: 0.2
Offer: 0.0

Jacob Bieber <j.bieber@bthx.com>
Tue 12/08/2022 09:20
To: e.smith@bthx.com

Hello,

We have received a request from the BTHX IT Security Office to review recent account activity associated with your account, specifically that your BTHX account had been used from several different locations in a short time frame, specifically Tokyo Metropolis.

There are several possibilities for this kind of behavior:

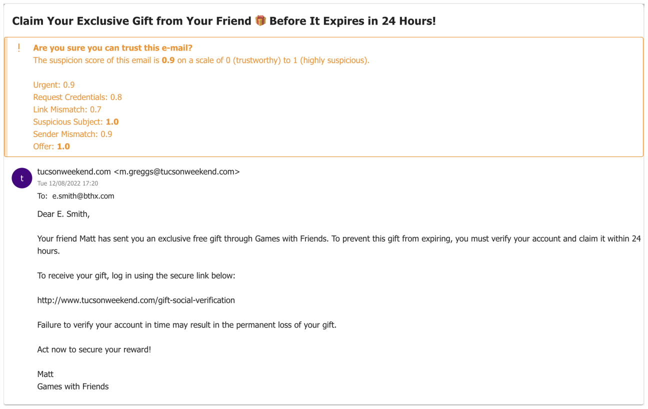
- your BTHX account is compromised
- password sharing with other users
- use of a third-party VPN

Can you please provide an explanation for this activity or let us know if you do not recognize this activity? If you are using a third party vpn, which vpn service are you using?

Sincerely,
Jacob Bieber

BTHX OIT
OIT Service Desk
Be Aware: security.oit.bthx
-WE will NEVER ask for your password-

Figure 12: Email 1 – Transparent Explanation



Claim Your Exclusive Gift from Your Friend 🎁 Before It Expires in 24 Hours!

! **Are you sure you can trust this e-mail?**
The suspicion score of this email is **0.9** on a scale of 0 (trustworthy) to 1 (highly suspicious).

Urgent: 0.9
Request Credentials: 0.8
Link Mismatch: 0.7
Suspicious Subject: **1.0**
Sender Mismatch: 0.9
Offer: **1.0**

tucsonweekend.com <m.greggs@tucsonweekend.com>
Tue 12/08/2022 17:20
To: e.smith@bthx.com

Dear E. Smith,

Your friend Matt has sent you an exclusive free gift through Games with Friends. To prevent this gift from expiring, you must verify your account and claim it within 24 hours.

To receive your gift, log in using the secure link below:

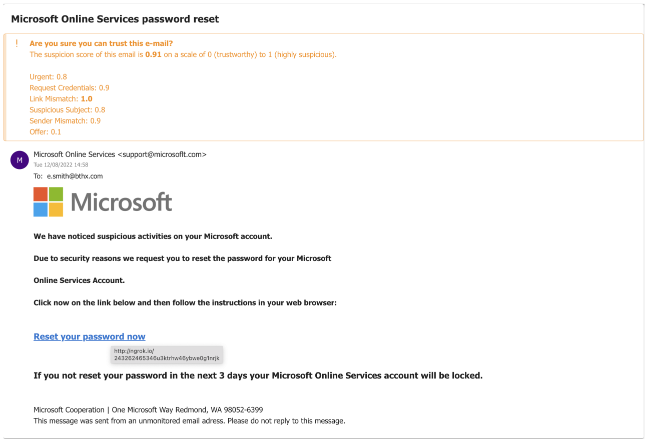
http://www.tucsonweekend.com/gift-social-verification

Failure to verify your account in time may result in the permanent loss of your gift.

Act now to secure your reward!

Matt
Games with Friends

Figure 13: Email 2 – Transparent Explanation



Microsoft Online Services password reset

! **Are you sure you can trust this e-mail?**
The suspicion score of this email is **0.91** on a scale of 0 (trustworthy) to 1 (highly suspicious).

Urgent: 0.8
Request Credentials: 0.9
Link Mismatch: **1.0**
Suspicious Subject: 0.8
Sender Mismatch: 0.9
Offer: 0.1

Microsoft Online Services <support@microsoflt.com>
Tue 12/08/2022 14:58
To: e.smith@bthx.com

Microsoft

We have noticed suspicious activities on your Microsoft account.

Due to security reasons we request you to reset the password for your Microsoft

Online Services Account.

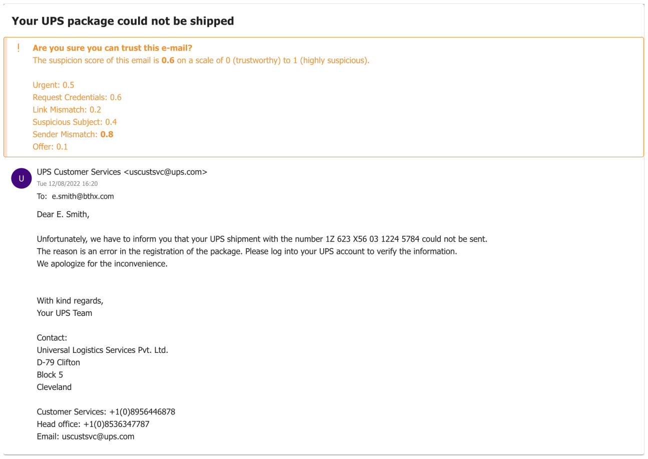
Click now on the link below and then follow the instructions in your web browser:

Reset your password now
http://ngrok.io/
243262465346u3ktrhw46ybwe0g1nrjk

If you not reset your password in the next 3 days your Microsoft Online Services account will be locked.

Microsoft Cooperation | One Microsoft Way Redmond, WA 98052-6399
This message was sent from an unmonitored email adress. Please do not reply to this message.

Figure 14: Email 3 – Transparent Explanation



Your UPS package could not be shipped

! **Are you sure you can trust this e-mail?**
The suspicion score of this email is **0.6** on a scale of 0 (trustworthy) to 1 (highly suspicious).

Urgent: 0.5
Request Credentials: 0.6
Link Mismatch: 0.2
Suspicious Subject: 0.4
Sender Mismatch: **0.8**
Offer: 0.1

UPS Customer Services <uscustsvc@ups.com>
Tue 12/08/2022 16:20
To: e.smith@bthx.com

Dear E. Smith,

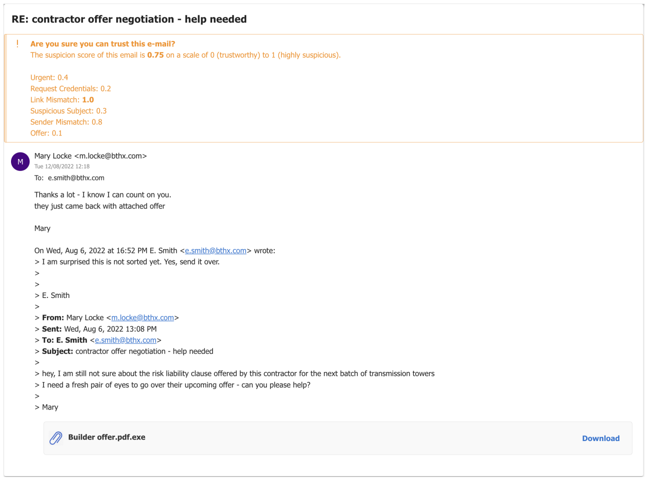
Unfortunately, we have to inform you that your UPS shipment with the number 1Z 623 X56 03 1224 5784 could not be sent.
The reason is an error in the registration of the package. Please log into your UPS account to verify the information.
We apologize for the inconvenience.

With kind regards,
Your UPS Team

Contact:
Universal Logistics Services Pvt. Ltd.
D-79 Clifton
Block 5
Cleveland

Customer Services: +1(0)8956446878
Head office: +1(0)8536347787
Email: uscustsvc@ups.com

Figure 15: Email 4 – Transparent Explanation



RE: contractor offer negotiation - help needed

! **Are you sure you can trust this e-mail?**
The suspicion score of this email is **0.75** on a scale of 0 (trustworthy) to 1 (highly suspicious).

Urgent: 0.4
Request Credentials: 0.2
Link Mismatch: **1.0**
Suspicious Subject: 0.3
Sender Mismatch: 0.8
Offer: 0.1

M Mary Locke <m.locke@bthx.com>
Tue 12/08/2022 12:18
To: e.smith@bthx.com

Thanks a lot - I know I can count on you.
they just came back with attached offer

Mary

On Wed, Aug 6, 2022 at 16:52 PM E. Smith <e.smith@bthx.com> wrote:
> I am surprised this is not sorted yet. Yes, send it over.
>
>
> E. Smith
>
> **From:** Mary Locke <m.locke@bthx.com>
> **Sent:** Wed, Aug 6, 2022 13:08 PM
> **To: E. Smith** <e.smith@bthx.com>
> **Subject:** contractor offer negotiation - help needed
>
> hey, I am still not sure about the risk liability clause offered by this contractor for the next batch of transmission towers
> I need a fresh pair of eyes to go over their upcoming offer - can you please help?
>
> Mary

Builder offer.pdf.exe **Download**

Figure 16: Email 5 – Transparent Explanation