Phishing Reporting in Organizations: What Motivates Employees to Take Action?

Pavlo Burda[1], Luca Allodi[2], Alexander Serebrenik[2], and Nicola Zannone[2]

[1]ICT Institute, Utrecht, The Netherlands (corr. author*)
[2]Eindhoven University of Technology, Eindhoven, The Netherlands

Abstract

Purpose - This study investigates the factors influencing employees' decisions to report suspicious phishing emails in organizations, addressing the gap in understanding what motivates users to report and which types of emails are most likely to be reported.

Design/Methodology/Approach - In this study, we sample and interview $n = 49$ employees from the pool of phishing reporters at a European university. Interviewees are selected based on the sophistication of the emails they report, considering both contextual and technical dimensions. We cluster reporters according to their (emerging) reporting behavior and conduct semi-structured interviews until thematic saturation is reached. Through thematic analysis, we identify 21 main themes that drive reporting.

Findings - The results indicate that the primary drivers for reporting suspicious emails are the desire to protect and help the organization and coworkers. Additional factors include a sense of responsibility, awareness of potential consequences, and feelings of insecurity. Participants are more likely to report phishing emails that appear well-impersonated and with a believable pretexts, signaling user prowess in estimating the potential impact of phishing attacks.

Originality/Value - This research offers a novel perspective on the complex interplay between motivations to report with a discussion in the broader theoretical context, as well as on the practical implications of our findings.

Keywords: Phishing, Reporting;

1 Introduction

Phishing attacks are a major threat to organizations and private citizens alike. Automated phishing detection and filtering are measures commonly in place in most organizations, yet phishing emails regularly pass those filters and end up in users' inboxes. At that point, the user is the last line of defense against an attack that, if successful, may pose risks to the entire organization. Cyber security awareness campaigns, training, and phishing simulations are generally aimed at improving users' ability to *detect* phishing attacks (Allodi et al., 2020); on the other hand, a single successful attack can allow an attacker to successfully breach, for example, through subsequent lateral movement attacks (Ho et al., 2019). Therefore, it is key that the organization is in a position to take swift action upon the arrival of a new attack. *Reporting* is the main mechanism on which organizations rely. It has received the attention of several recent research contributions (Jensen et al., 2017; Kwak et al., 2020; Lain et al., 2022; Burda et al., 2020a; Kersten et al., 2022; Marin et al., 2023; Distler, 2023). User reporting is a mechanism that allows users to report suspicious emails to central analysis units (sometimes hosted within the organization, if large enough, or outsourced to service providers) that can then take action on that information. This may include updating automated detection filters, blocking associated domains, and/or checking (e.g., through the employment of a monitoring infrastructure such as a Security Operation Center) whether reported emails led to other users clicking on suspicious links or opening malicious attachments. Unlike user reporting in general, phishing reporting is currently a relatively understudied topic. Crucially, users are the main driver behind a successful defense mechanism, at least in part relying on reporting (Burda et al., 2020a; Chen et al., 2024).

The important factors reported driving users' decision to report phishing emails include the technical confidence they have in using the mechanism (*self-efficacy*) (Kwak et al., 2020; Marin

*p.burda@mailbox.org

et al., 2023), the characteristics of the attacks (Burda et al., 2020a; Distler, 2023), and knowledge of organizational policies on phishing reporting (Marin et al., 2023). Interestingly, the role of the organization appears to emerge frequently in the literature as associated with a user's likelihood to report. For example, keeping the reporter informed about the outcome of their report appears to motivate users to continue reporting (Jensen et al., 2017; Williams et al., 2018; Lain et al., 2022). Similarly, previous research showed the role of personality traits associated with the so-called *positive cybersecurity behaviors*, within the broader context of *Organizational Citizenship Behaviors* (OCB), in affecting a user's propensity to report phishing (Marin et al., 2023). Indeed, reporting assumes the traits of discretionary behavior that individuals engage with to the benefit of the organization as a whole without specific obligations or rewards to do so. In this context, it becomes critical to understand what *motivates* individuals in reporting phishing emails in an organizational setting. Doing so will allow us to design better policies, instruments and processes to identify and act on reported phishing and create safer IT environments in which users can operate (including those not inclined to report) (Burda et al., 2024). Importantly, the link between reporting behavior and motivations to report has not yet been fully explored, leaving a large gap in the characterization of security behavior and open questions on how best to nudge or motivate users towards positive cybersecurity behaviors (Burda et al., 2024; Marin et al., 2023). To address this gap, this work builds on and extends our previous work (Author(s), 2024) and answers the following question.

What are the factors driving employee decisions to report suspicious phishing emails?

To answer this question, we collaborated with the Operations Security Team of our institution, a medium-large technical university in Europe, to analyze 8369 emails reported by employees over 766 days. We first analyzed the reported emails to identify the general emerging reporting behavior of individual reporters in terms of the reporting frequency and the type of reported emails. For the latter, we devised heuristics to evaluate at scale to what degree a reported email is contextually and technically sophisticated with respect to the organization's environment. For example, an email spoofing our institution's domain would be considered technically sophisticated. Similarly, an email mimicking internal communication styles would be considered contextually sophisticated. Evaluating the characteristics of the emails reported through lenses of technical and contextual sophistication, we clustered reporters based on their *emergent behavior* in reporting: e.g., users that tend to report highly sophisticated emails on both dimensions are more likely to be clustered together than not. We then employed these clusters as 'strata' to sample employees and interviewed them to gauge what motivates them in reporting. We iteratively coded interview transcripts to identify key emerging themes and continued sampling from said clusters until we reached 'thematic saturation', i.e., no new themes emerged from the last two interviews in that cluster. We reached thematic saturation after $n = 49$ interviews with as many employees.

Our contribution is multi-faced. We identify a number of themes that motivate employees to report. The interplay between these themes is complex and ranges from the desire to protect the organization and help less security-conscious colleagues to the desire to fight back and neutralize attackers. These can be used to shape organizational policies on reporting, as well as awareness programs focusing on outcomes that align well with users' motivations. Awareness, doubt, and (technical) self-efficacy play an important role in determining whether an employee will report a suspicious email. Interestingly, 'doubt' can be a motivating factor pushing employees to report 'just in case' the email might be malicious. Well-impersonated phishing emails were judged as the most dangerous and urgent to report, signaling user prowess in estimating the potential impact of highly believable attacks. Participants mostly relied on sender cues and pretext to identify inconsistencies and motivate their reporting preferences where an unconvincing sender with a weak pretext induce a lesser intention to report . We discuss our findings in the broader theoretical context of Protection Motivation Theory (PMT), identifying several parallels between identified themes and PMT, thus suggesting that it may represent a meaningful framework for evaluating phishing reporting mechanisms.

Outline The paper is structured as follows. Section 2 discusses the relevant background. Section 3 presents our methodology and data. Section 4 and Section 5 respectively present the results and discuss our findings. Section 6 concludes the paper.

2 Background and Related Work

2.1 Phishing reporting

Phishing reporting plays a critical role in an organization's cybersecurity response strategy whereby employees who detect phishing attempts may notify the relevant IT department of an ongoing campaign. Effective reporting facilitates the initiation of remediation procedures, such as blocking access to malicious domains or notifying employees who might still be at risk of falling victim to the attack. This rapid response is crucial, as research shows that a significant portion of targeted employees are victimized within the first few hours of the attack's deployment (Burda et al., 2020a). In response to the growing threat of more sophisticated phishing attacks, such as spear phishing, research has increasingly focused on phishing reporting and the factors that influence it (Stembert et al., 2015; Jensen et al., 2017; Kwak et al., 2020; Lain et al., 2022; Burda et al., 2020a; Kersten et al., 2022; Marin et al., 2023; Burda et al., 2023; Distler, 2023). Previous work explored how organizations can improve reporting, for example, by examining incentives to report (Jensen et al., 2017), by identifying 'naturally immune' individuals (Burda et al., 2020a), and testing reporting effectiveness in the field (Lain et al., 2022). Among factors influencing reporting, most of the research explored individual factors, such as perceptions, beliefs, and attitudes toward the organization (Kwak et al., 2020; Marin et al., 2023), and contextual factors, such as user interface, job role, and situation (Stembert et al., 2015; Lain et al., 2022; Distler, 2023). However, what reasons and motivations drive individuals to report phishing attacks in organizations remains unclear. For instance, Burda et al. (2020a) and Distler (2023) interviewed 12 and 14 reporters of a spear phishing campaign, respectively. The findings reveal that employees may be unable to generalize the rationale for reporting a suspicious email, stating various reasons for reporting, such as being aware of the sophistication of an attack or feeling responsibility, and *not* reporting due to ill-perceived liability or lacking efficacy towards phishing. In another study with nine participants, Burda et al. (2023) suggest that, in addition to feeling responsible for colleagues, some employees may decide to only report phishing emails that are more sophisticated or more *believable* than 'generic' phishing. However, evidence suggests that more believable emails are less likely to be reported because, in principle, less detectable (Kersten et al., 2022). In general, phishing reporting depends on a variety of motivations, attitudes, and the type of phishing emails encountered, and as such, it can be considered an emergent behavior arising from the interactions of these factors.

2.2 Phishing believability

We build on prior research on phishing sophistication and reporting to define *phishing believability*– the extent to which a recipient perceives a phishing email as a credible message from the claimed source (Kersten et al., 2022). Higher believability reduces the likelihood of phishing emails being detected and reported to the IT department. Several factors influence phishing believability. These include technical details, such as the sender's address or the payload URL (Jakobsson et al., 2007; Dhamija et al., 2006; Parsons et al., 2015b; Molinaro and Bolton, 2018), and context alignment, which refers to how well the email matches the target's context, such as impersonation or premise (Steves et al., 2020; Burda et al., 2020b). Other aspects such as persuasion techniques (Van Der Heijden and Allodi, 2019; Valecha et al., 2022) and the overall appearance (Zielinska et al., 2016; Williams and Polage, 2019) play a critical role in effectively convincing recipients of the email's authenticity.

Our focus is on the factors that are commonly found in phishing emails and that enhance 'technical believability' and 'contextual believability', as these are critical to understanding why phishing attempts often succeed. Specifically, *Technical believability* refers to the convincing or realistic nature of the email from a technical viewpoint (Steves et al., 2020; Kersten et al., 2022). This comprises spoofing of the sender address, hiding the URL payload behind legitimate services, such as well-known URL shorteners or hosting services, using homograph techniques to mask the sender or URL domains, or disguising email attachments. *Contextual believability* concerns the alignment with the context or expectations of the recipient (Greene et al., 2018; Kersten et al., 2022). This includes tailoring the contents of the email to the targets, such as impersonating the target's organization or related entities, or crafting details that seem relevant to the recipient's environment, such as recent activities or typical coming and going interactions in the workplace. Table I describes the email features often used to improve technical and contextual believability.

Table I: Mail believability features. Source: Authors own work.

	Feature	Description
Technical believability	*Payload URL*	The link requiring target interaction should be plausible or difficult to distinguish from a genuine one Molinaro and Bolton (2018); Steves et al. (2020).
	Sender	Non-trivial, coherent spoofing or masquerading of Sender name and address (username and/or domain) Parsons et al. (2015a); Steves et al. (2020).
	Attachment	The type of document attached to the email does not appear to be malicious or obviously harmful Steves et al. (2020). The name of the attachment indicates a harmless file Dewan et al. (2014).
	Other headers	Other email headers, such as To: and Reply-To:, can be spoofed as part of a phishing tactic Sethuraman et al. (2024).
Contextual believability	*Impersonation*	The manipulation of email features in such a way the email seems to originate from a legitimate source. For example, mentioning or purporting to be the target organization in the sender, subject, or body of the email Dhamija et al. (2006); Jakobsson et al. (2007); Parsons et al. (2015b).
	Premise alignment	The email content aligns with target experiences and expectations, such as references to events or activities happening in the target's environment, or aligning the pretext with typical internal procedures or jargon at an organization Pastor-Galindo et al. (2020); Steves et al. (2020).
	Timing	The delivery of the attack at a useful time, e.g, festivities or busy hours Binks (2019).

2.3 Research gap and contribution

Phishing reporting in organizations depends on various factors, such as employees' attitudes, job roles, and the type of phishing email encountered (Kwak et al., 2020; Stembert et al., 2015; Marin et al., 2023), as well as their motivations (Burda et al., 2020a, 2023; Distler, 2023). However, research on why employees report phishing is limited, often based on small participant samples and simulated campaigns. This indicates a lack of comprehensive understanding of how employees decide to report phishing.

To address this gap, research must identify emerging behaviors in phishing reporting and the factors that influence them, such as what employees choose to report. For example, the perceived sophistication and danger of a phishing email can impact reporting decisions (Burda et al., 2020a, 2023). Previous studies suggest that highly sophisticated phishing emails are reported less often than generic ones (Kersten et al., 2022). In this study, we aim to better understand reporting motivations by interviewing 49 employees from a mid-sized European university who reported phishing emails between 2019 and 2021.

3 Methodology

Since our research question is exploratory in nature, suitable research methods include those that offer rich, qualitative data about a phenomenon, allowing building hypotheses and tentative theories (Easterbrook et al., 2008). To this end, we adopted semi-structured interviews (Bird, 2016) to collect data. The advantage of interviews over, e.g., surveys, is the possibility to ask follow-up questions and clarify respondents' answers.

To sample participants, we analyzed a dataset of emails reported by employees to the IT department of a mid-sized European university (referred to as *UNI*) (Section 3.1). At the time of the study, UNI's reporting process required employees to forward suspicious emails to a shared inbox called the *abuse inbox*, monitored by IT security staff. The IT team used these reports to detect threats and take actions such as domain or IP blocking. After completing investigations, the team provided feedback to reporters about the findings and any actions taken. We note that, at the time of this study, UNI did not conduct official awareness campaigns or training aimed at improving phishing detection by users or influencing their reporting behavior.

We gained authorized access to UNI's *abuse inbox* for our analysis, allowing us to identify employees with similar reporting patterns based on the contextual and technical characteristics of the emails they reported. Using these insights, we applied stratified sampling (Baltes and Ralph, 2022) to select employees representing different reporting behaviors for interviews (Section 3.2). During the interviews, we explored their motivations for reporting and asked them to evaluate four phishing emails in terms of how likely they would report them (Section 3.3). We used the initial stage of the Socio-Technical Grounded Theory (STGT) method (Hoda, 2022) to analyze the qualitative data. STGT involves collecting and analyzing data through theoretical sampling and

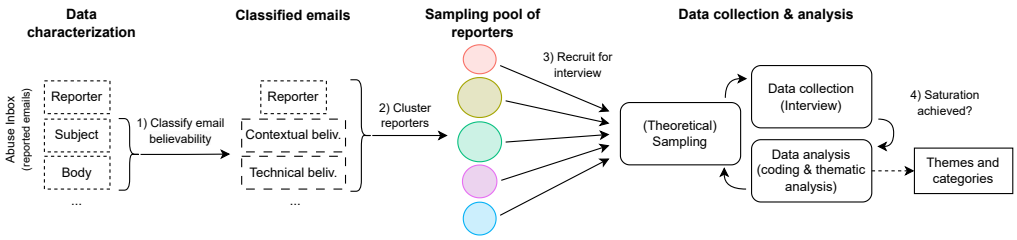


Figure 1: Methodology overview. Source: Authors own work.

Table II: Abuse inbox variables and counts. Source: Authors own work.

Variable	Count
id	8369
reporterAddress	1460
toAddress	1921
fromAddress	3178
subject	3119
body	6503
attachmentName	687
attachment	1118
receivedTime	5102
reportedTime	8356

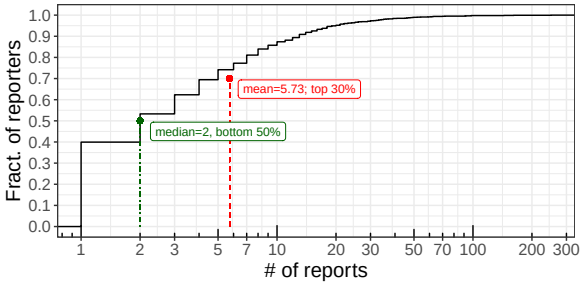


Figure 2: Fraction of reporters and # of reports (log10 scaled). Source: Authors own work.

thematic analysis. A summary of the approach is shown in Fig. 1.

3.1 Data characterization

To understand the distribution of reports and reporters, we first characterized the dataset of emails reported to the UNI IT department (hereafter *abuse inbox*). This step aims to determine employees with similar emergent behaviors (cf. Section 3.2). The dataset extracted from the abuse inbox at UNI contains 8369 individual emails spanning from 2019-02-08 to 2021-03-15 for a total of 766 days. Due to the location of UNI, the dataset includes emails in both Dutch and English.[1]

A summary of the dataset is given in Table II (a detailed summary is available in the Appendix A). All variables but `id` contain duplicate values, e.g., the same emails from a phishing campaign can be reported by multiple employees. The variable `reporterAddress` represents the 1460 unique reporters. The distribution of the number of reports per reporter is skewed towards a small number of reports per reporter (Mdn = 2, mean = 5.73), with the top 25% of employees reporting at least 6 emails (Q3 = 6). Fig. 2 presents the fraction of reporters per number of reports.

3.2 Sampling and recruitment

Theoretical sampling ensures an adequate sample size when no new insights emerge after repeated cycles of data collection and analysis, a point known as thematic saturation (Hoda, 2022). Following established guidelines (Francis et al., 2010; Guest et al., 2020), we started with an initial group of participants and iteratively expanded the sample based on thematic analysis outcomes until reaching saturation, where no additional insights were identified. Using random sampling directly from the pool of reporters could lead to thematic saturation before capturing behaviors associated with less common cases, such as employees reporting rare or highly sophisticated phishing emails (Hoda, 2022). To address this, we adopted stratified sampling. First, we categorized emails based on features related to contextual and technical believability (Section 3.2.1). Next, we clustered employees based on the types of emails they reported, including those with high believability (Section 3.2.2). Finally, we selected participants from each cluster for interviews (Section 3.2.3).

3.2.1 Email classification

We first randomly sampled 20 reporters and manually reviewed 131 emails reported by them to determine which believability features (cf. Section 2.2) can be used to automatically classify emails. To ensure the quality of the review (McDonald et al., 2019), four investigators applied the criteria

[1]One of the authors is a proficient Dutch speaker. The automated processing described in Section 3.2.1 accounts for the language of the reported email.

defined in Table I, discussed each sampled email, identified and resolved points of disagreements, and iteratively updated the review of emails. Given the limited amount of emails to be coded, two co-coding sessions were enough to resolve disagreements and update the criteria definitions.

After excluding reports of non-phishing emails[2], the review suggests that the sampled emails vary largely in content and mostly present a low believability on either the contextual dimension, technical dimension, or both. Emails deemed highly believable on both dimensions are relatively rare (≈14%, see Appendix A in the supplementary material). Importantly, we observe a high variability of features affecting contextual and technical believability (i.e., `fromAddress`, `subject`, and `body`). This makes the implementation of automated solutions for accurately labeling reported emails hard or impossible, as it would require manually labeling thousands of email features by trained agents with contextual knowledge of UNI's environment. However, to build our employee sampling pool, it suffices to have an *approximate* representation of email believability. We assigned each email a binary value, *high* or *low*, indicating the presence or absence of contextual believability and technical believability features. Stemming from Table I, the feature matching rules used for the classification are detailed in Table XI in Appendix A. The classifier achieves an accuracy of 88.6% and a precision of 76.1% for contextual believability, and an accuracy of 72.5% and a precision of 78.1% for technical believability (see Appendix A for further details).

We find that emails with low technical and contextual believability are the most common in our dataset (4805, 57%). Emails with high technical believability are the least common, equal split between high (745, 9%) and low contextual believability (750, 9%). High contextual believability and low technical believability are relatively common (2069, 25%).

3.2.2 Clustering of reporters

The sheer majority of reporters in our dataset reported suspicious emails only once or twice (Mdn = 2). As our goal is to understand users' motivation to decide to report or not report a suspicious email, we are interested in identifying subjects that do show repeated reporting behavior. The average number of reports per person is 5.7, corresponding to the top 30% of the reporters. Hence, we chose five reports as a threshold for prospective interviewees. This results in a pool of 445 subjects. As we are interested in individuals' behavior and decision-making, we filtered out reporters whose email address corresponds to a shared functional account, such as 'library@UNI.edu', because it is unfeasible to trace back individuals who reported a specific message in the past from a given shared address. Among 445 subjects, 336 used a personal, as opposed to functional, email address.

Each reporter can be characterized in terms of the classes of emails they have reported, i.e., each reporter can be described using four variables: `Ch_Th` (fraction of high contextual and high technical believability emails among all emails they have reported), `Cl_Th` (contextual low and technical high), `Ch_Tl` (contextual high and technical low) and `Cl_Tl` (contextual low and technical low). Using these four variables as a representation of reports, we perform clustering. We do not expect clearly separated groups as it is likely that almost all reporters have reported the most prevalent type of phishing at some point (i.e., contextual low and technical low). To determine a suitable number of clusters, we applied the elbow method (Thorndike, 1953) and chose five clusters (see Fig. 6 in Appendix A). Since the features (i.e., `Ch_Th`, `Cl_Th`, `Ch_Tl` and `Cl_Tl`) are numeric and at the same scale (i.e., bounded between 0 and 1), to perform the clustering we applied the Hartigan and Wong k-means algorithm (Morissette and Chartier, 2013) with 25 random sets for the initialization and the Euclidean distance.[3]

Fig. 3 shows a visualization (reduced to two dimensions) of the five clusters, capturing approximately 70% of the overall variance. Cluster 1 overlaps with Clusters 2, 3, and 5. Other clusters appear disjoint over these two dimensions; given the high fraction of captured variance, this suggests the clustering succeeded in meaningfully separating users based on emergent reporting behavior. Clusters 1, 3, and 4 have approximately 45 subjects each, whereas clusters 2 and 5 have 99 subjects (cf. Table III). Clusters 2 and 5 include users reporting mostly (i.e., approx 80% of the time) 'technical low' emails (`Ch_Tl` or `Cl_Tl`). The other three clusters show a prevalence of 'high believability' emails over either technical, contextual, or both dimensions.

3.2.3 Recruitment of participants

The recruitment was performed in several iterations following the theoretical sampling approach. We send invitations to the institutional email address the prospective interviewees used for report-

[2]Not phishing: 42 out of the 131 considered reported emails (32%), of which 28 are spam emails and 14 are legitimate emails.

[3]The algorithm is implemented in the R package `stats`.

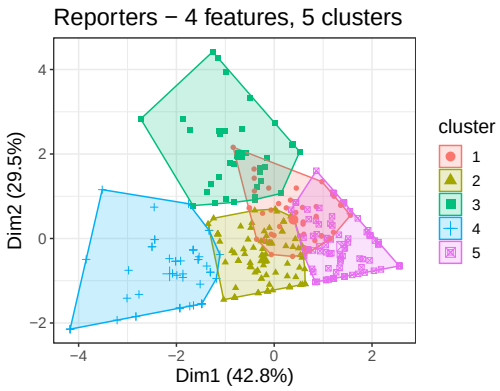


Figure 3: Reporter clusters projected on the first two dimensions of principal component analysis of the email features explaining 69.4% of variance. See Table III for the total number of reporters in each cluster and the mean values of `Ch_Th`, `Cl_Th`, `Ch_Tl` and `Cl_Tl` per cluster. Source: Authors own work.

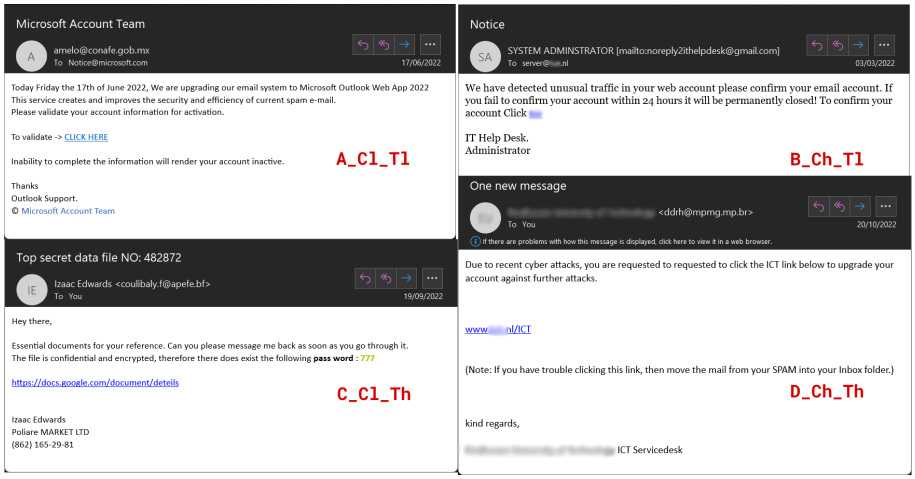


Figure 4: Shown phishing emails with varying believability: `A_Cl_Tl`, `B_Ch_Tl`, `C_Cl_Th`, `D_Ch_Th`. Source: Authors own work.

ing to the abuse inbox. From the 336 reporters in the five clusters, we randomly sample subjects from each cluster. To keep the scheduling of interviews manageable, for each iteration, we sample no more than 10% of the cluster size.

Following the theoretical sampling procedure (Hoda, 2022), we invited participants until thematic saturation was reached *for each cluster separately*. Based on Guest et al. (2020), we deem saturation to be achieved when no new themes emerge from the analysis of the last two interviews. We reached thematic saturation after four rounds, and 49 interviews (cf. Table III).

3.3 Interviews

In the first part of the interview (Q1), we asked participants high-level questions about their motivations for reporting suspicious emails to the abuse inbox. In the second part (Q2), we asked them to rank four phishing emails based on how likely they were to report each one to IT, and to explain their reasoning. The emails, shown in Figure 4, were selected from the abuse inbox based on believability criteria and displayed within the Outlook application used at the institution. Following best interview practices (Bird, 2016), we tested and revised a one-page interview guide after a pilot interview with a PhD student in our research group. A detailed interview guide is available in Appendix C. The semi-structured interviews, as recommended by Bird (Bird, 2016), were conducted by one of the investigators in a confidential setting, either in-person (in individual offices or reserved conference rooms) or online via the institutional video conferencing service. Before the interview, participants were reminded they could withdraw from the study without explanation and were provided with an information sheet and consent form. The primary language was English, except for one case where Dutch was used, with a proficient

Table III: Clusters of reporters. C=contextual and T=technical, l=low and h=high. Source: Authors own work.

Cluster	Ch_Th	Cl_Th	Ch_Tl	Cl_Tl	Total	Invited	Interviewed
1	0.054	0.270	0.161	0.515	46	20	10
2	0.075	0.042	0.336	0.547	99	16	12
3	0.385	0.074	0.162	0.380	44	23	8
4	0.068	0.075	0.613	0.244	48	27	10
5	0.059	0.061	0.103	0.777	99	27	9
Total					336	90	49

Dutch speaker ensuring the validity of the interview questions and answers.

3.4 Data analysis

The interviews were recorded, transcribed with the help of specialized software, and paraphrased following the approach taken in a similar study (Distler, 2023). The subsequent analysis of the transcripts is largely based on the initial stage of the STGT method interleaving data collection and data analysis steps: 1) we identified emerging codes, 2) formed and assessed themes, 3) checked the saturation of themes, and 4) eventually sampled additional participants from our sampling pool (see Section 3.2). The first iteration of coding was carried out in person with hard-copy printouts by four investigators (one of them with extensive experience with qualitative methods). The following iterations were carried out by the first author in a virtual environment. To ensure the quality of the analysis, all four investigators regularly met and discussed the coded transcripts, identified ambiguities, and resolved disagreements. Furthermore, to ensure the validity of the codes, the remaining authors have independently applied the codes created by the first author to three randomly selected paraphrased transcripts each. The inter-rater agreement (McHugh, 2012) of the first author with the second one was 80%, with the third one was 68%, and with the fourth one was 77%, for an average agreement of 75% (a 'moderate' agreement level (McHugh, 2012)). More details on the codes are available in Appendix D.

3.5 Ethical considerations

Data collection and analysis were carried out under ethical approval [ANONYMIZED] by our institution's ethical review board. Potentially sensitive information willingly or unwillingly contained in the reported messages was removed from the analysis whenever possible. Participants were thoroughly informed about the research aims and methods, both orally and in writing, on a consent form; they were offered enough time to familiarize themselves with the form and ask further questions (including after the interview). The interview was carried out adhering to ethical guidelines (University of Chicago, 2020). Data collection and analysis were executed on the university's premises, through encrypted communication and storage, minimizing potential harm to the participants. Based on the consent agreement signed by the participants, we cannot publish the raw data.

4 Results

4.1 Rationale for reporting phishing emails (Q1)

4.1.1 Thematic analysis

We identified 13 themes characterizing the rationale for reporting phishing emails to the IT department by the employees (Q1). Table IV summarizes the identified themes and participant contributions per cluster. We stress that the counts of participants merely reflect the opinions of the interviewees participating in our study and are not intended to reflect the prevalence of the themes among all reporters. As such, these numbers are intended to provide qualitative rather than quantitative insights.

The table shows that the intention to protect their colleagues, the organization, and themselves was mentioned most commonly by the participants. The second most common reason is helping their colleagues or the organization as a whole. Following, various, less homogeneous motivations drive employees' reporting behavior. For example, some employees report in case of doubt (ask for confirmation), others are aware of phishing risks due to their previous experience (awareness

Table IV: Overview of the themes identified from the answers to Q1 per cluster of reporters. The numbers of contributing participants are not intended to reflect the prevalence of the themes among reporters from UNI. Source: Authors own work.

Theme	**Contributing participants**					
	Tot	**clst1**	**clst2**	**clst3**	**clst4**	**clst5**
`Protect others`: colleague or community, their systems, confidentiality and data	34	6	9	6	7	6
`Protect UNI`: organization, employer, their systems network and data	24	4	6	5	3	6
`Protect me: myself and my system, confidentiality and data`	18	5	4	3	2	4
`Help UNI/IT`: intention to assist the university, the IT department in handling the issue of phishing	29	5	8	5	5	6
`Help others`: intention to assist colleagues in handling the issue of phishing	3	1	1	0	1	0
`Loyalty`: to the organization and community as a reason to protect	1	1	0	0	0	0
`Awareness & experience`: awareness of phishing risks and consequences stemming from personal and third party experience	8	2	1	2	2	1
`Ask for confirmation`: reporting in case of doubt over the legitimacy of an email to IT or colleagues	6	2	0	1	1	2
`Sense of responsibility`: the feeling of responsibility or duty to report as a norm, civic duty, or conscientious behavior.	10	3	3	0	2	2
`Efficacy to report`: knowing and being confident about how or why to report	7	2	1	1	3	0
`Annoyance`: the feelings of being annoyed or angry by the unwanted emails or the sender	8	0	4	1	1	2
`Fight hackers`: the desire to fight or punish the attackers, or a feeling of disdain towards the perpetrator as a reason to report	5	2	1	1	1	0
`Detection`: reporting because an email was detected as (suspicious) phishing	5	2	0	0	0	3

`and experience`). Less common reasons include `annoyance` and the will to '`fight hackers`'. Interestingly, participants mention `detection`, a distinct but related activity to reporting, as a motivation. In the following, we elaborate in detail on the identified themes. Appendix D provides further examples in the codebook.

Protect and help. The `protect others` theme reflects the desire to prevent colleagues from falling victim to phishing or its consequences, such as causing *"harm to our computers"* (P2). This was the most common reason for reporting, mentioned by almost 70% of participants. Many participants were concerned that others might not recognize phishing emails and could be deceived or *"be hacked"* (P7). These concerns were often based on their own experiences with phishing (see `awareness and experience`). A few participants thought their colleagues lacked the skills to detect phishing, with one noting, *"[I report] so that nobody gets trapped in those phishing emails. I don't, but others can"* (P1). However, one participant observed that anyone could be deceived under certain circumstances. Some participants were also concerned about colleagues being annoyed by phishing emails: *"Prevent other people from being annoyed by yet another phishing email"* (P44).

The second most common reason for reporting is to `protect UNI`, which focuses on safeguarding the university as an institution rather than individual colleagues. Participants aimed to reduce institutional risk, with reasons such as, *"to lower risk for the organization"* (P24) and *"to protect the confidentiality at UNI"* (P13), or to avoid reputation damage, as in *"to keep the reputation of UNI high"* (P41).

The `protect me` theme represents motivations related to safeguarding participants' own security, including data, work, and the ability to work. Some participants recognized that protecting others is linked to protecting themselves, as one noted: *"Maybe one day I'm a bit sleepy and I press the wrong link, so I'd like to prevent others as well because it can happen to anyone"* (P44).

The `help UNI`, which includes helping the IT department, reflects the desire to support the organization in maintaining a secure environment. Participants mentioned reporting to aid the IT department by providing visibility into ongoing attacks: *"I report to make the IT department aware, so they can act on it, and they can't monitor everything"* (P9) or *"My motivation is to let [the IT department] investigate and have more clues about the phishing emails"* (P45). Some participants linked this motivation to helping others or themselves: *"I think they [the IT department] may be able to block links etc., so sending it might help others and avoid them [phishing emails]"* (P34).

Awareness, experience, and doubt. We observe a connection between the themes of protecting and helping and the theme `experience and awareness`. For example, participants referred to

news about data leaks and ransomware attacks at hospitals and universities, with the University of Maastricht being frequently mentioned. These references show an awareness of how phishing attacks can happen and their potential consequences. Participants shared insights such as, *"Relative to previous years' attacks, current attacks are way more realistic, like no spelling mistakes"* (P45) and *"I know the danger phishing can cause, e.g., in Maastricht"* (P9).

At least three participants referenced past incidents, suggesting that awareness of risks may come from experience. One participant explained, *"Some colleagues were affected and their computers were locked and held hostage (crypto ransomed), with their PCs unavailable for two weeks. Therefore, the impact can be big"* (P39).

When in doubt about an email's legitimacy, participants often used a `ask for confirmation` strategy as a "default", which is a common recommendation in organizations and awareness materials. However, it may conflict with efficient security processes, as too many reports can overwhelm IT teams and delay responses to higher-priority cases (Burda et al., 2020a). For example, one participant mentioned asking colleagues before reporting to the IT department: *"I also ask for confirmation from my colleagues if they received it too. If so, they often tell me to 'Throw it away' and don't report"* (P29). Another participant described reporting as their default approach: *"Always, if I don't trust, I send it to 'abuse' [the reporting inbox]"* (P47). This default strategy is closely linked to the `efficacy to report` theme, which involves knowing how and what to report.

Efficacy and responsibility. The `efficacy to report` theme captures mentions of employees' confidence in reporting suspicious emails, focusing on the act of reporting rather than the ability to identify phishing. Most participants indicated they report because the IT department advises them to, reflecting a "default strategy" from the `ask for confirmation` theme. As one participant explained, *"I just send them to [the IT department] and I get a response later that this is indeed phishing and that I don't have to do anything"* (P34). This behavior is supported by knowing how to report, as highlighted by statements such as *"I was told if I get a suspicious email, I should report, so I did"* (P34) or *"Because I know that there is the abuse [inbox]"* (P19). Another participant mentioned the ease of reporting, noting it takes *"very little effort"* (P23), which further encourages action.

The `sense of responsibility` theme reflects participants' motivation to report out of duty and concern for others. One employee stated, *"If someone threatens via my email, I think it's my duty to inform [the IT department]"* (P11), while another compared reporting to *"throwing garbage in a civil way"* (P48). Some participants linked their responsibility to the collective security of the organization, as expressed in *"Security is our all responsibility"* (P32). This sense of conscientiousness often stems from a broader commitment to the organization as a community, aligning with the `loyalty` theme. For example, one participant emphasized their responsibility for *"the data in the emails"* (P4), which they viewed as part of their role as a secretary.

Annoyance and fight hackers. Many participants reported phishing emails due to the `annoyance` caused by unwanted messages. Some expressed anger towards the sender, with comments such as *"Do they think I'm stupid, or something?"* (P15) and *"Get a life, a job, do something better"* (P44). Others viewed phishing emails as just another form of junk mail: *"If this is spam, it annoys me; therefore, I reported some"* (P31). However, there was a clear distinction between phishing and spam, as one participant noted, *"It's pretty rare to receive [phishing] emails, so it's a very low level of being annoyed"* (P20).

In at least two cases, `annoyance` was linked to a desire to `fight hackers`. Participants expressed disdain for hackers, with statements like *"I hate them"* (P9) or *"They should stay out of it"* (P38). One participant mentioned the need to prevent hackers from profiting: *"To minimize the chance to enrich themselves by our means"* (P26). This suggests that reporting phishing emails is driven not only by logical reasons but also by emotional responses (Burda et al., 2023).

Detection. The `detection` theme is a special case of reasons to report, as answers belonging to this theme somewhat elude the scope of reporting, which is an action that typically follows detection. For example, participants stated that they report emails *"When I get an email with a strange request or the email address is not correct"*, (P5) or *"When I don't believe it, it's too good to be true."* (P9), indicating that "they report because they detect". It is worth noting that all respondents in this theme provided additional reasons, such as to `protect` or `help UNI`. Yet, the immediate answer to why they do report suspicious emails was related to detection. This can be a sign of unbalanced prowess between detection and reporting activities (Burda et al., 2020a).

4.1.2 Differences across clusters

We observe some differences between clusters 2, 3, and 4, and clusters 1 and 5. Clusters 1 and 5 are the only ones that mention `detection`-related reasons for reporting phishing emails. Participants in these clusters also reported more contextual low and technical low (`Cl_Tl`) emails than the other types (see Table III). This suggests that their reasoning of "report everything detected" aligns with their behavior of reporting any suspicious email.

A notable difference is in cluster 5, where the `ask for confirmation` theme is prominent as a "default strategy" for handling suspicious emails. This may indicate that participants in cluster 5, who mostly report `Cl_Tl` emails, are less confident in assessing phishing emails and prefer to report all suspicious emails. This behavior could be due to a lack of confidence in detecting phishing or a preference for following organizational rules conscientiously.

4.2 Email judgments (Q2)

4.2.1 Thematic analysis

Recall that the interviewees have been asked to rank the emails of Figure 4 according to the likelihood of them reporting the email to IT over the other emails. For example, a participant is more likely to report email `D_Ch_Th` over `B_Ch_Tl`, `B_Ch_Tl` over `A_Cl_Tl`, and `A_Cl_Tl` and `C_Cl_Th` ranked equally last— we record this preference `D_Ch_Th`>`B_Ch_Tl`>`A_Cl_Tl` = `C_Cl_Th`. The unique rankings made by participants and their frequency are shown in Table XVI (Appendix D). By employing the ranked pairs voting system to identify the ranking winners (Tideman, 1987), we obtain that `D_Ch_Th` defeats `A_Cl_Tl`, `B_Ch_Tl` and `C_Cl_Th`; `B_Ch_Tl` defeats `A_Cl_Tl` and `C_Cl_Th` and, finally, `A_Cl_Tl` defeats only `C_Cl_Th`. In other words, `D_Ch_Th` is the preferred email to report to IT by the participants, and `C_Cl_Th` is the least preferred. However, a significant amount of participants deviate from this pattern. For example, a notable case is that four participants rank the emails on the same level, where they would report all four of them. In some cases, `C_Cl_Th` and `A_Cl_Tl` (with the lowest ranking) are ranked first to report, signaling a high variability in user reasoning on what makes an email worth reporting.

To get further insights into what influences user ranking preferences, we asked participants to reason over the email characteristics and their ranking decisions. We identified nine themes that characterize employees' preferences over reporting or not reporting the four emails shown during the ranking task. Three of the themes are further divided into subthemes (six subthemes in total) that provide further structure to employees' reasoning. Table V reports the nine themes and six subthemes with counts of participants' mentions of each theme.

Relevance cues. Table V shows that the most cited email features belong to `Relevance cues` theme (92). This theme appears as one of the main motivations for the high ranking of `D_Ch_Th`, in particular, `D_Ch_Th` being highly targeted at UNI. The subtheme `Targetization` captures the user focus on the relevance of the organization-targeted nature of the emails. For example, P1 clearly explained that "*`D_Ch_Th` is first, because it wants to give the impression it's from the university: the link is UNI but hovering, it's outside university, especially the domain (not UNI.nl).*". Also, P33 mentioned "*[`D_Ch_Th` first] because it says UNI ICT Service desk and the sender name is UNI, but if you look a bit more you see a different sender address*". The targeting theme can be linked to the `Protect others/UNI` of Section 4.1 as a reason to report suspicious phishing emails. In this sense, also *absence* of `Targetization` can be used to judge an email as not worth reporting: "*`C_Cl_Th` is not targeted [...] and less urgent to my opinion. It looks like malware-related. I hope that our anti-malware solutions would prevent further [damage]*" (P37).

The `Relevance cues` theme concerns the relevance dimension of an email, and as such, it is at the cross-way of the `Sender cues` and `Pretext danger perceptions` themes. Specifically, the subtheme `Mismatch between sender, content or links` encompasses all mentions of inconsistency between emails' features and participants' expectations. One common inconsistency relates to the mismatch of the sender's name or address: "*I'd report `C_Cl_Th` because it's an unknown Sender and the Sender name doesn't match the email address username.*"(P16). Other mentioned inconsistencies concern the link – "*`D_Ch_Th` is not from the university and if I click [hover] on this link, then I see it's not from the university, then it's super suspicious.*" (P1) – or even the content – "*`B_Ch_Tl` and `A_Cl_Tl` are equally first because [...] 'email upgrade' is not from IT but from Microsoft (which is unlikely)*". (P2). The mismatch theme illustrates a primary strategy used by the participants to rank the emails and is a constant appearance for all four emails and for the majority of participants (39 mentions in Table V). For example, given the high contextual believability of `D_Ch_Th` and `B_Ch_Tl`, several employees point out the inconsistency between the impersonated elements – "*`D_Ch_Th` is more likely [to report] because, if the email address is*

Table V: Overview of the identified themes in Q2. The count of contributing participants are not intended to reflect the prevalence of the themes among reporters from UNI. Source: Authors own work.

Theme	Sub-theme	Description	Contributing participants					
			Tot	clst1	clst2	clst3	clst4	clst5
Relevance cues		Mentions of cues or features related to the relevance dimension of the email features or contents.	92	23	26	20	13	10
	Targetization	The evaluation of emails' targetization of organization or related entities.	49	7	13	15	8	6
	Mismatch between sender, content or links.	An inconsistency between email features is perceived.	39	15	11	4	5	4
	Premise alignment	The email's content aligns with the target's experiences and expectations.	4	1	2	1	0	0
Sender cues		Cues related to the sender (name, address) or features related to the sender.	53	18	12	4	8	11
	Sender untrustworthy	The sender appears either unknown or untrustworthy.	37	13	8	1	4	11
	Sender disguise as trustworthy or legitimate	The sender features impersonate or disguise as trustworthy.	16	5	4	3	4	0
Pretext danger perceptions		Mentions of danger perceptions of the pretext (the semantics of the email features and contents).	48	12	10	9	8	9
Persuasion cues		The email pretext employs persuasion techniques, such as urgency or fear.	14	6	2	3	1	2
Content cues		Mentions of email cues or features that relate to the content of the email body.	23	4	12	2	2	3
	Content has links	The email body has a link or pretext that nudges to follow the link.	15	2	9	2	1	1
Email etiquette		Mentions of emails features related to the overall look&feel (e.g., tone, syntax, spacing, font, etc.).	19	3	6	4	1	5
Obvious phishing		Not reporting because the email is considered obvious phishing, SPAM or unwanted, with no further distinction.	27	6	8	5	8	0
Others might fall for it		Reporting because the co-workers might be deceived by the email.	13	4	6	0	1	2
Report all		Reporting all email without specific reasons (applies to all four emails).	9	4	3	1	0	1

obviously wrong, the name says UNI, the link says UNI.nl, so I consider this to be more likely to be clicked. Also, because, it really references the university." (P26); and `B_Ch_Tl` – "*[...] it says SYSTEM ADMINISTRATOR, if I get such an email, I expect something from IT, also in that case I check the email address for what I expect and the link inside to be from the UNI domain.*" (P1). As with `Targetization`, `Mismatch between sender, content or links` includes user reasoning over placing an email low in the rank, such as ranking `C_Cl_Th` low " *[...] because we never send document with google docs with passwords, I think is not okay.*" (P14).

An interesting case of relevance is the `Premise alignment` sub-theme which includes mentions of consistency between the email's premise and user context or experience. Some participants deviate from the most common ranking of `D_Ch_Th` 'high' and `C_Cl_Th` 'low' by acknowledging the plausible premise disguise of `C_Cl_Th` and `A_Cl_Tl` and considering them more dangerous. For example, one participant prefers `A_Cl_Tl` because it "*refers to something that is going to happen at UNI. But it depends on the situation, when the attacker is smart and knows what about the context here. Something that is plausible in this context.*" (P40). Another uncommon preference concerns `C_Cl_Th`: "*I use a lot google docs (on my private email). If you only look at the content, you might think that it's legit, but this email address is strange.*" (P2). The relevance of contextual elements such as the premise may sway some participants preferences as it strongly depends on user context.

Sender cues. `Sender cues` have been recognized across all four emails as `Sender untrustworthy` and `Sender disguise as trustworthy and legitimate` themes. Inspecting the sender appears to be another important strategy that employees utilize to rank emails and potentially report them (53 mentions in Table V). For instance, "*They all have strange email address (e.g., from Brazil, Gmail account)*" (P38) and "*[I report] when I see a link related to UNI or ICT. Or when someone is trying to spoof somebody else.*" (P15). These judgments apply to all four emails. While many participants cut short their judgment by saying that the sender is untrustworthy ("*[...] sender address has .br, this is dodgy.*" (P21)) or inconsistent ("*the email address is (immediately) fake,*

or not UNI", P12), some participants drew their attention to the impersonated nature of the sender disguised as trustworthy or legitimate: "*D_Ch_Th here they are indicating UNI, even if there is a different email address*" (P31).

Similarly to the previous themes, some participants mention sender features when deciding to rank the email low (no intention of reporting it): "*A_Cl_Tl everyone can see that the sender address is untrustworthy; I might report it but can also tell that everyone can recognize that this is stupid.*" (P28). This observation highlights how users may develop their reasoning over reporting in opposite directions while starting from the same basic email cues.

Pretext and Persuasion. The `Pretext danger perceptions` theme covers a significant number of participant judgments over the dangers of the email pretext (48). As with the targetization subtheme, employees' `Pretext danger perceptions` bring `D_Ch_Th` high in the ranking due to its pretext related to 'recent cyber attacks' and 'the need to upgrade user account' (Figure 4). This is a topical matter at UNI and many participants mentioned the danger of this pretext: "*[...] the text about cyber attacks looks important to us.*" (P28) or "*D_Ch_Th is very misleading*" (P10). We observed the same focus on pretext's danger for `B_Ch_Tl` which also ranks high in Table XII: "*B_Ch_Tl seems the most plausible, the most proper email and the most dangerous one.*" (P3) and "*B_Ch_Tl looks like generic, but is targeted at UNI (custom build for UNI)*" (P37). While `A_Cl_Tl` and `C_Cl_Th` are not explicitly targeted at UNI (both have contextual believability low, Table I), several participants recognized the danger of pretext and targetization in both, e.g., "*A_Cl_Tl says "Microsoft Account Team" and we all use Microsoft stuff here, Outlook. That might be a trigger, it's obviously so fake, but to think it's more authentic because is related to products that we use, and of course upgrades (pretext) is something that happens periodically.*" (P26).

The apparently dangerous pretext of `D_Ch_Th` or `B_Ch_Tl` did not, however, make all the participants rank them as more urgent to report. Instead, a participant explains that `D_Ch_Th` is "*much more friendly* (P27)" than the other emails, P34 says "*B_Ch_Tl there is not really anything in it: no explanation in the message and it's obviously spam. I'd probably not report it.* and P13 "*B_Ch_Tl is a bit generic and badly crafted email, so maybe easier to spot*". The participants used such judgments of `Pretext danger perceptions`, that is, the bad craft and generic nature of the pretext, to lower the rank. Even more interestingly, participant P2 acknowledges that they might not be able to recognize `D_Ch_Th` as phishing, and thus rank it low, because "*we had recently a similar communication. [...] and we already received a communication that this was phishing training*", showing that the contextual situation for `D_Ch_Th`'s pretext can make a substantial difference in employees' perceptions of danger.

The perceived danger (and the consequent high ranking) has been reported by participants also in terms of fear and urgency appeals, which fall under the `Persuasion cues`. Specifically, `B_Ch_Tl` (and somewhat `A_Cl_Tl`) was consistently mentioned to convey urgency – "*B_Ch_Tl and A_Cl_Tl are equally first because of the urgency pretext*" (P2). Another participant reports "*'Inability to complete the information will render your account inactive' is a bit threatening [in A_Cl_Tl]*" (P27). Together with UNI-targeted mail, persuasion signals drove `B_Ch_Tl` into a high-rank position, revealing a certain degree of user awareness with regard to the typical (and still effective) techniques used in phishing attacks.

Content cues and Email etiquette. On top of sender and pretext cues, employees regarded as important various `Content cues`, most notably when the `Content has links`: " *[...] the 'to validate' or 'click here', this kind of links I don't really like them.*" (P11). The more prominent the link was in the email, as in `A_Cl_Tl`, the more participants tended to mention it in their judgments. For example, P2 "*get[s] suspicious with large 'click here'*" and for P15 "*the 'click here' [A_Cl_Tl] makes is easy for somebody to actually click.*".

Also the looks and feeling of the overall email features, `Email etiquette`, was a common reference in many cases. For instance, the presence of typos ("*[A_Cl_Tl has] typos, it looks more deceiving*", P10), an informal salutation ("*C_Cl_Th starts with 'Hey there', that's for me very suspicious, nobody start's emails like that.*", P8) or the formatting "*the colors of the password text [in C_Cl_Th] and the sender address... [are suspicious]*" (P12) brought some participants to place `C_Cl_Th` and `A_Cl_Tl` higher in the rank.

Differently from the previous themes, however, content cues and email etiquette were the most common starting point to judge an email lower in the reporting ranking. For example, due to an odd subject: "*the subject [A_Cl_Tl] "Microsoft Account Team" it's words that we don't use; then it would say "Microsoft Outlook" or something. [...] I'd just delete it.*" (P7); or the looks: "*A_Cl_Tl looks like any invite email from a company to follow a course or seminar or a LinkedIn invite.*" (P11). `C_Cl_Th` was especially penalized with the sheer majority of mentions of content and email etiquette cues used to discount `C_Cl_Th` to the lower ranks, thus, underlining the weight of the

email content and appearance in user judgments. One participant quoted: "*here you would smile because C_Cl_Th has a 'password'. You can be triggered because the URL is spelled bad. [...] Even if it has "https" in the URL, it doesn't matter.*" (P4).

Report all, Others might fall for it and Obvious phishing. Some participants ranked all four emails on the same level (see Table XII in Appendix D) and consistently mentioned that they would undoubtedly report all four emails irrespective of their characteristics. `Report all` captures such cases whereby, for example, P3 motivates their judgment comparing it to "*[...] a management test, 'Do you like herrings or apple crumble?'. I don't know, I would report them all.*". Another participant mentioned that they are unable to judge and therefore would report all of them, akin to the `Ask for confirmation` theme of Section 4.1: "*I can't judge the amount of harm that they can do. I'm sure they can all do harm, but I can't judge, and I won't.*" (P6).

Participants that mentioned their preference for ranking an email given the danger of a co-worker falling for it, were included in the `Others might fall for it` theme. Similarly to the report-all case, this theme directly connects to `Protect others` of Section 4.1, for example "*[...] a person not 'from IT' can think that [A_Cl_Tl] is legitimate.*" (P20). However, many employees justified their preference for giving a low rank to emails due to their appearance as `Obvious phishing` or spam. For instance, participant P19 disregards `A_Cl_Tl` `B_Ch_Tl` and `C_Cl_Th`: "*I don't report this, because it's really clear that it's phishing (not good enough) and we get these all the time*". Interestingly, even `D_Ch_Th` was mentioned to be an obvious phishing attack: "*D_Ch_Th is a little bit recognizable*" (P39).

4.2.2 Differences across clusters

Similarly to Section 4.1.2, there are limited differences between clusters in terms of preferences for certain emails or themes. In terms of ranking, there is no difference across clusters: by computing the ranking winners with the ranked pair voting system, in all five clusters the preference winner is `D_Ch_Th` and the loser is `C_Cl_Th`. However, looking at Table V, we notice that certain clusters of participants provided more mentions of specific themes and subthemes: Clusters 2 and 3 emphasize the `Targetization` theme, Cluster 2 mentioned `Content cues` much more often, and Clusters 1 and 2 cited `Sender` and `Mismatch of sender cues` more frequently than the rest. Participants in Clusters 2 and 3 reported more contextual high (`Ch`) emails than those in Clusters 1 and 5 (see Table III). This might explain the higher number of mentions of `Targetization`. However, this observation is at odds with Cluster 4, which reported high contextual emails more than any other cluster, but refers to `Targetization` less often. More interestingly, no one from Cluster 5 judged an email as `obvious phishing` or cited the `sender disguise` theme, but rather preferred to mention the `untrustworthiness of sender features`. As Cluster 5 reports mostly `Cl_Tl` emails, a possible explanation could be due to a lower confidence with respect to reporting activities that we observed in Section 4.1.

5 Discussion

5.1 Lessons Learned

Summary of results. In Q1, all participants mentioned protecting others/UNI or assisting the organization as a reason to report, making these themes the main factors driving the reporting of suspicious emails. These factors can be directly linked to the altruistic tendency of individuals in organizations (an antecedent of OCBs), which is a known predictor of intention to report phishing (Marin et al., 2023). The remaining themes vary considerably in presence between participants and are often related to threat appraisals (Wang et al., 2017), self-efficacy (Kwak et al., 2020) or emotional drivers (Burda et al., 2023), underlining the high dimensionality of factors that constitute the rationale for reporting. For Q2, the ranking showed a clear preference for reporting `D_Ch_Th` due to its targeted features and dangerous pretext. Targeted phishing increases both reporting rates and susceptibility (Burda et al., 2020b), while pretexts can either enhance or weaken their effectiveness (Greene et al., 2018). Participants judged emails based on sender and content appearance, focusing on inconsistencies. `C_Cl_Th` faced the most criticism for its unconvincing appearance and weak pretext, making it less likely to be reported.

A complex mix of motivations to report. Our results for Q1 suggest that while some themes, such as those related to `helping` and `sense of responsibility`, are common among most reporters, participants report a wide variety of additional motivations. These motivations

often overlap, such as in the themes of `helping` and `sense of responsibility`, where some individuals feel both an obligation (e.g., from internal policy) and a desire to be proactive and helpful within their organization and towards their colleagues. Interestingly, *uncertainty* plays a role in reporting, with some participants acting "just in case" a threat is present (`ask for confirmation`). We also identified new motivations for reporting phishing emails: `annoyance` and `fight hackers`. `Annoyance` is an intrinsic motivation to get rid of unwanted emails, distinguishing phishing from spam, while `fight hackers` signals a more personal, visceral motivation to act. Since these motivations have not been explored before and are not linked to existing models, they could be tested in future research to better understand reporting behaviors. We did not observe notable differences across clusters (possible explanations are discussed in Section 5.3). Qualitative observations from clusters 1 and 5 (reporting mostly low-context and low-technical believability emails) suggest that motivations to report in these clusters are mainly linked to `ask for confirmation` and `detection`. As discussed in Section 4.1.2, this may indicate a lower ability to handle suspicious phishing emails in clusters 1 and 5 compared to cluster 3, which mostly reports high-believability emails. If this relationship holds, a phishing reporting mechanism could provide "reasons to report" options in the email client's interface (e.g., similar to (Stembert et al., 2015)), including an `ask for confirmation` option. This could benefit users who repeatedly select `ask for confirmation` by offering targeted training to improve their phishing detection skills (Chen et al., 2024). While this study focuses on motivations for reporting phishing emails, future research could explore different motivations for reporting other types of suspicious emails, such as spam. Although these may not involve actual security threats, studying them could reveal additional insights into how threat perception influences user actions.

The dangers of targetization and pretext in user judgments. The email ranking task (Q2) revealed a clear preference for reporting `D_Ch_Th` and `B_Ch_Tl`, while `A_Cl_Tl` and `C_Cl_Th` were seen as less worthy of reporting. Contextual believability (`Ch`) appeared more influential than technical soundness (`Th`). For instance, despite `C_Cl_Th`'s high technical believability, it was ranked lowest due to the weak perceived danger of its pretext. As previous studies show, a pretext that is not salient or plausible is often ignored and perceived as harmless (Williams and Polage, 2019; Steves et al., 2020). Participants consistently favored `D_Ch_Th` for its targeted features, `relevance cues`, and `dangerous pretext`, aligning with findings on the effectiveness of spear phishing (Bullee et al., 2017; Burda et al., 2020b; Distler, 2023). Similarly, `B_Ch_Tl` was highly rated due to targetization and the perceived danger of `persuasion cues`, such as urgency and authority, which are known to increase phishing susceptibility (Williams et al., 2018; Parsons et al., 2019; Van Der Heijden and Allodi, 2019). Across all emails, participants frequently relied on the `mismatch between sender, content, or links` to identify inconsistencies. Interestingly, at least six participants ranked `C_Cl_Th` and `A_Cl_Tl` higher than expected, with some even ranking `D_Ch_Th` and `B_Ch_Tl` lower. This was often driven by `premise alignment`, where the emails' content strongly matched users' contexts, overriding other features. Contextual relevance remains a key factor in phishing susceptibility, though its variability makes it difficult to predict overall susceptibility (Greene et al., 2018; Distler, 2023; Burda et al., 2024; Sommestad and Karlzén, 2024).

The use of `sender cues` by many participants (in conjunction with the mismatch of sender features) remains a positive signal of employees relying on the most effective email cue to recognize phishing attacks (Parsons et al., 2015a; Molinaro and Bolton, 2018). Therefore, our findings on the prevalence of sender cues in participants' evaluations support the idea that effective security training should rely on few, but effective cues such as sender features (Steves et al., 2020). Still, some participants mentioned `content cues` and `email etiquette` of the email as a way to evaluate email legitimacy. This goes against the best security practices whereby email legitimacy should not be judged primarily on the email looks, as attackers are steadily improving in crafting authentic-looking artifacts (Mossano et al., 2020; Butavicius et al., 2022). This observation indicates that there is still room for improvement in anti-phishing education: for example, the longstanding recommendation to 'look for poor spelling and grammar' is falling short, for example, in the view of attackers employing AI-based services to scale up the quality of their craft (Schmitt and Flechais, 2024).

Some participants follow along their motivations to report phishing, such as `ask for confirmation` and to `protect others`, deciding to `report all` the emails without distinctions or by associating the phishing email features to the risk of `others falling for it`. On the other hand, a substantial number of employees have a clear idea of what emails, i.e., `obvious phishing`, should *not* be reported to IT. For example, P31's motivation to not report is "*when it's not useful for UNI to know about it.*" and P11 comments that "*we (employees) should sort ourselves which email is interesting or not. Because otherwise IT gets 500 emails a day*" suggesting that indeed too many

reports may hinder the effectiveness of the IT team at UNI.

The importance of keeping the reporter 'in the loop'. Another promising research venue on reporting concerns how feedback and (public) acknowledgment of employees successfully reporting attacks might benefit organizational security. Previous work showed that acknowledging reported incidents and validating reported emails can facilitate reporting as a crowd-sourced defense (Jensen et al., 2017; Williams et al., 2018; Lain et al., 2022). In this sense, our findings lead to ask the question: given employees' motivations and reasons, what feedback could be provided to encourage reporting again in the future? For example, one interviewee stated: *"Sometimes, when reporting, there is a lack of feedback [...] I might lose motivation because nobody reads this [the report]."* (P13). Indeed, personal feedback on true positive reports encourages employees to report more (Lain et al., 2022). Moreover, P13 added *"It would be interesting to have internal statistics once a year or so, maybe this will inspire people to report more"*. This suggests that, by reporting public statistics on recent phishing campaigns and reporting efforts, employees whose main concern is protecting colleagues might consolidate their self-efficacy as well as motivate others (Jensen et al., 2017, 2022; Marin et al., 2023). However, apart unintended effects from incentivized reporting (Burda et al., 2020a; Jensen et al., 2022), not everyone cares to receive feedback and might instead perceive it as a nuisance: *"When I report an email to [the IT department], they open a ticket and inform me what they've done with this, but I don't care, I'm not curious what they do with it. I just want to give them information."* (P7). Therefore, the question of the *type* of feedback to give reporters assumes a new relevance on its own.

Links to theoretical underpinnings. Our findings reveal a range of motivations behind phishing email reporting, with many participants expressing a desire to protect their organization or help others. These motivations reflect both a sense of personal responsibility and a concern for organizational well-being, which align with theoretical frameworks such as Protection Motivation Theory (PMT). PMT explains how individuals assess threats and coping strategies when deciding whether to engage in protective behaviors, such as reporting phishing emails (Rogers, 1975; Sommestad et al., 2015). It posits that behaviors result from two appraisal processes: threat appraisal (e.g., perceived severity and vulnerability) and coping appraisal (e.g., self-efficacy and response costs). Several themes emerging from our data resonate with these components. For example, `sense of responsibility` and `efficacy to report` reflect coping appraisals, such as a belief in one's ability to take effective action. Similarly, `awareness and experience` and `pretext danger perceptions` suggest heightened threat appraisals; for instance, participants recognize phishing as a serious risk and feel capable of addressing it. While PMT has frequently been used to model avoidance behaviors in cybersecurity, such as not clicking on suspicious links, its application to the proactive behavior of phishing reporting remains underexplored (Marin et al., 2023). Our findings suggest that PMT could provide a valuable lens for understanding not just why people avoid threats, but also why some take the extra step to report them. Importantly, this perspective may be particularly useful for identifying why some individuals choose not to report. Future research could build on our findings by applying PMT to model barriers to reporting and design interventions—such as targeted training or awareness materials—that enhance perceived efficacy or reduce perceived effort in reporting (Chen et al., 2024).

Similar to PMT, many themes reflect how traits influencing Organizational Citizenship Behaviors (OCBs) relate to phishing reporting. For example, the distinction between protecting colleagues and the organization can be framed as individual- vs. organization-directed OCBs (Organ, 1997; Williams and Anderson, 1991). Participants who focus on protecting the organization may share traits related to organization-directed OCBs, while those protecting others may score higher on individual-level OCB traits. For example, altruism, a predictor of intention to report (Marin et al., 2023), aligns with the themes of `protecting` and `helping`. Other organization-level OCB traits, such as commitment and civic virtue, appear linked to reporting behaviors motivated by `protect UNI` and `sense of responsibility`, although these connections have not been confirmed in previous work (Marin et al., 2023). Our findings related to the `ask for confirmation` also raise questions about whether more reports lead to better security, as excessive reporting may overwhelm IT teams (Burda et al., 2020a). Some individuals might avoid reporting `obvious phishing` to reduce the burden on colleagues, a behavior aligned with the sportsmanship OCB trait (Marin et al., 2023). Further research into the drivers of phishing reporting is necessary to develop a sustainable security culture (Marin et al., 2023) and improve collective defenses against phishing (Burda et al., 2020a; Chen et al., 2024).

5.2 Implications for Practitioners

This study offers actionable insights for organizations seeking to improve phishing email reporting and develop more effective, inclusive awareness strategies. Our analysis reveals that altruistic motivations, such as protecting colleagues and supporting the organization, are central drivers among frequent reporters. These motivations align with concepts from Protection Motivation Theory and Organizational Citizenship Behaviors, suggesting that reporting can be framed as a form of corporate citizenship. Organizations can cultivate a sense of shared responsibility by emphasizing how individual reports of suspicious emails help protect colleagues and support organizational security. Notably, these motivations emerged despite the absence of formal training or consistent awareness material at the organization where the study was conducted. Our findings suggest that reporting behavior was shaped by individual values, prior experience, or general security intuition rather than by coordinated institutional efforts. As such, awareness campaigns may benefit from explicitly reinforcing these naturally occurring pro-social motivations, while also targeting users who are less engaged. Personalized strategies that appeal to different motivational profiles, such as self-efficacy, concern for others, or uncertainty, could encourage broader participation across the organization.

Our findings also underscore the need to support users with varying levels of confidence and detection ability. Many participants reported emails "just in case", highlighting the need for tools that support uncertain users. For instance, reporting interfaces could include an optional "reason for reporting" field to help security teams interpret user intent and tailor feedback or training accordingly. At the same time, concerns about overburdening IT teams led some participants to avoid reporting "obvious phishing", highlighting the importance of clear and consistent guidance on what should and should not be reported, along with scalable feedback mechanisms. While some users appreciated personalized acknowledgment or updates, others preferred minimal follow-up. Scalable solutions, such as periodic summaries of phishing campaigns or anonymized statistics, could offer a practical balance, reinforcing engagement without overwhelming IT resources. Additionally, our results indicate that users still rely on superficial cues such as email appearance to assess legitimacy. Awareness efforts should therefore emphasize high-impact, resilient indicators, such as sender information and content inconsistencies, while moving away from outdated heuristics such as spotting grammatical errors, which become increasingly unreliable in the age of tailored, AI-assisted phishing (Allodi et al., 2020).

Although this study focused on frequent reporters to understand consistent engagement, the findings also point toward the need for future research exploring the perspectives of infrequent or non-reporters. Understanding their motivations and barriers could clarify why awareness efforts reach only a subset of users and inform more inclusive interventions. In this sense, this study provides a foundation for targeted, evidence-based strategies to strengthen organizational resilience against phishing. By aligning training, reporting tools, and feedback mechanisms with user motivations and behaviors, organizations can foster more sustainable and responsive security cultures.

5.3 Limitations/Threats to validity

As is the case with any empirical study, the validity of our conclusions might be threatened by various reasons. Our research combines a quantitative (identification of the clusters) and a qualitative component (analysis of interviews). For the quantitative component we adhere to the well-established threats to validity framework of Shadish, Cook, and Campbell (Shadish et al., 2002) and guidelines based on it (Wohlin et al., 2012).[4] This framework is inappropriate for the qualitative component, and hence in our reflection, we adhere to the guidelines of Lincoln and Guba (Lincoln and Guba, 1985). **Quantitative.** One of the main *constructs* of our study is believability; the validity of our conclusions can, hence, be threatened by the operationalization of this construct. These threats are partly inherited from the previous work on this topic (Greene et al., 2018; Kersten et al., 2022) and partly stem from the joint identification of the believability features suited for automation (Section 3.2.1). To address the latter threats, we have to ensure that agreement has been reached between all four investigators. This threat is closely related to the instrumentation, i.e., errors introduced by the believability classifier (Section 3.2.1), that threaten the *internal validity* of the study. To mitigate this, the development of the classifier followed best

[4]We are mindful of the ongoing debate on trade-offs in research study design and threats to validity induced by these trade-offs, taking place in many scientific disciplines (Capano and Engeli, 2022; Mercer et al., 2007; Wolff and Haase, 2020). However, in the absence of a commonly accepted alternative to the threats to validity framework of Shadish, Cook, and Campbell, we adhere to it, acknowledging its limitations (Reichardt, 2011; Robillard et al., 2024).

practices involving iteratively analyzing false positive and false negative outcomes over independent training sets until classification performance was deemed sufficient. Further, UNI conducts internal phishing awareness campaigns that may affect the reporting rates considered by our sampling strategy. As the degree to which awareness campaigns may affect behavior is unknown, we rely on our internal knowledge of UNI and employ a best-effort approach and remove sampled reported emails that are likely to belong to an internal awareness campaign. We expect any residual effect to be minimal and not impacting study results. *External validity* related to the criteria used to classify emails that may depend on the type or frequency of emails normally received at UNI (e.g., whose employees may be used to emails from, for example, predatory publishers). **Qualitative.** To ensure *credibility* of the study, we focus on UNI, a university we have been familiar with for an extended period. To ensure *transferability* of our findings, in this report, we provide a detailed description of the interview and analysis process such that the person who might be interested in transferring the study insights can decide whether they might apply to their situation. Finally, to ensure *dependability*, i.e., the ability to audit the process, we provide the audit trail from the interview data to process notes (themes and examples of quotes corresponding to each theme).

Finally, we discuss the **limitations** of our work. The exploratory nature of the study required us to focus on the qualitative analysis and no quantitative insights should be derived from it. The overview of the themes per cluster in Tables IV and V does not necessarily reflect the prevalence of different reasons for reporting. Follow-up studies should investigate several complementary explanations for the lack of differences across clusters: a strong variability of motivations (other than protecting) between subjects, fundamentally different base rates of the type of phishing emails received by participants, and the classification of phishing believability is not able to reflect the real believability of emails to derive the clusters.

6 Conclusion

In this work, we investigated what motivations drive users to report suspicious emails and which types of emails they report. To this end, we sampled and interviewed $n = 49$ employees from the pool of phishing reporters at a medium-sized European technical university. The results show that protecting and helping the organization and others are the main factors that drive the reporting of suspicious emails. Other factors for reporting phishing emails are a sense of responsibility, awareness of the negative consequences they can lead to, or insecurity. Interestingly, our results show phishing reporting is also driven by feelings such as annoyance and anger towards the attackers. Employees judged phishing emails by sender and content appearance, preferring to report targeted emails with a convincing pretext, and discounting inconsistent and unappealing emails. By relating our findings with PMT, we identified relevant insights and promising directions for future work.

References

L. Allodi, T. Chotza, E. Panina, and N. Zannone. The Need for New Antiphishing Measures Against Spear-Phishing Attacks. *IEEE Security & Privacy*, 18(2):23–34, 2020.

S. Baltes and P. Ralph. Sampling in software engineering research: a critical review and guidelines. *Empirical Softw. Engg.*, 27(4), 2022.

A. Binks. The art of phishing: past, present and future. *Computer Fraud & Security*, 2019(4): 9–11, 2019.

C. Bird. Interviews. In *Perspectives on Data Science for Software Engineering*, pages 125–131. Morgan Kaufmann, 2016.

J.-W. Bullee, L. Montoya, M. Junger, and P. Hartel. Spear phishing in organisations explained. *Information & Computer Security*, 25(5):593–613, 2017.

P. Burda, L. Allodi, and N. Zannone. Don't Forget the Human: a Crowdsourced Approach to Automate Response and Containment Against Spear Phishing Attacks. In *European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 471–476. IEEE, 2020a.

P. Burda, T. Chotza, L. Allodi, and N. Zannone. Testing the Effectiveness of Tailored Phishing Techniques in Industry and Academia: A Field Experiment. In *International Conference on Availability, Reliability and Security*, ARES '20, pages 1–10. ACM, 2020b.

P. Burda, L. Allodi, A. M. Altawekji, and N. Zannone. The Peculiar Case of Tailored Phishing against SMEs: Detection and Collective Defense Mechanisms at a Small IT Company. In *European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 232–243. IEEE, 2023.

P. Burda, L. Allodi, and N. Zannone. Cognition in social engineering empirical research: A systematic literature review. *ACM Trans. Comput. Hum. Interact.*, 31(2):19:1–19:55, 2024.

M. Butavicius, R. Taib, and S. J. Han. Why people keep falling for phishing scams: The effects of time pressure and deception cues on the detection of phishing emails. *Computers & Security*, 123:102937, 2022.

G. Capano and I. Engeli. Using instrument typologies in comparative research: Conceptual and methodological trade-offs. *Journal of Comparative Policy Analysis: Research and Practice*, 24 (2):99–116, 2022.

X. Chen, M. Sacré, G. Lenzini, S. Greiff, V. Distler, and A. Sergeeva. The Effects of Group Discussion and Role-playing Training on Self-efficacy, Support-seeking, and Reporting Phishing Emails: Evidence from a Mixed-design Experiment, 2024. arXiv:2402.11862 [cs].

P. Dewan, A. Kashyap, and P. Kumaraguru. Analyzing social and stylometric features to identify spear phishing emails. In *Symposium on Electronic Crime Research*, pages 4–8. IEEE, 2014.

R. Dhamija, J. Tygar, and M. Hearst. Why Phishing Works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 581–590. ACM, 2006.

V. Distler. The Influence of Context on Response to Spear-Phishing Attacks: an In-Situ Deception Study. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–18. ACM, 2023.

S. Easterbrook, J. Singer, M.-A. Storey, and D. Damian. Selecting Empirical Methods for Software Engineering Research. In *Guide to Advanced Empirical Software Engineering*, pages 285–311. Springer London, 2008.

A. Ferreira, L. Coventry, and G. Lenzini. Principles of Persuasion in Social Engineering and Their Use in Phishing. In *Human Aspects of Information Security, Privacy, and Trust*, LNCS, pages 36–47. Springer, 2015.

J. J. Francis, M. Johnston, C. Robertson, L. Glidewell, V. Entwistle, M. P. Eccles, and J. M. Grimshaw. What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology & Health*, 25(10):1229–1245, 2010.

K. Greene, M. P. Steves, M. F. Theofanos, and J. A. Kostick. User Context: An Explanatory Variable in Phishing Susceptibility. In *Network and Distributed Systems Security (NDSS) Symposium*. Internet Society, 2018.

G. Guest, E. Namey, and M. Chen. A simple method to assess and report thematic saturation in qualitative research. *PLOS ONE*, 15(5):e0232076, 2020.

G. Ho, A. Cidon, L. Gavish, M. Schweighauser, V. Paxson, S. Savage, G. Voelker, and D. Wagner. Detecting and Characterizing Lateral Phishing at Scale. In *USENIX Security Symposium*, pages 1273–1290. USENIX Association, 2019.

R. Hoda. Socio-Technical Grounded Theory for Software Engineering. *IEEE Transactions on Software Engineering*, 48(10):3808–3832, 2022.

M. Jakobsson, A. Tsow, A. Shah, E. Blevis, and Y.-K. Lim. What Instills Trust? A Qualitative Study of Phishing. In *Financial Cryptography and Data Security*, volume 4886 of *LNCS*, pages 356–361. Springer, 2007.

M. Jensen, A. Durcikova, and R. Wright. Combating Phishing Attacks: A Knowledge Management Approach. In *International Conference on System Sciences*, page 10, 2017.

M. L. Jensen, R. T. Wright, A. Durcikova, and S. Karumbaiah. Improving Phishing Reporting Using Security Gamification. *Journal of Management Information Systems*, 39(3):793–823, 2022.

L. Kersten, P. Burda, L. Allodi, and N. Zannone. Investigating the Effect of Phishing Believability on Phishing Reporting. In *European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 117–128. IEEE, 2022.

Y. Kwak, S. Lee, A. Damiano, and A. Vishwanath. Why do users not report spear phishing emails? *Telematics and Informatics*, 48:101343, 2020.

D. Lain, K. Kostiainen, and S. Čapkun. Phishing in Organizations: Findings from a Large-Scale and Long-Term Study. In *Symposium on Security and Privacy*, pages 842–859. IEEE, 2022.

Y. S. Lincoln and E. G. Guba. *Naturalistic Inquiry*. Sage Publications, 1985.

I. A. Marin, P. Burda, N. Zannone, and L. Allodi. The Influence of Human Factors on the Intention to Report Phishing Emails. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–18. ACM, 2023.

N. McDonald, S. Schoenebeck, and A. Forte. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):72:1–72:23, 2019.

M. L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282, 2012.

S. L. Mercer, B. J. DeVinney, L. J. Fine, L. W. Green, and D. Dougherty. Study designs for effectiveness and translation research: Identifying trade-offs. *American Journal of Preventive Medicine*, 33(2):139–154, 2007.

K. Molinaro and M. Bolton. Evaluating the applicability of the double system lens model to the analysis of phishing email judgments. *Computers & Security*, 77:128–137, 2018.

L. Morissette and S. Chartier. The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, 9(1):15–24, 2013.

M. Mossano, K. Vaniea, L. Aldag, R. Düzgün, P. Mayer, and M. Volkamer. Analysis of publicly available anti-phishing webpages: contradicting information, lack of concrete advice and very narrow attack vector. In *European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 130–139, 2020.

D. Organ. Organizational Citizenship Behavior: It's Construct Clean-Up Time. *Human Performance*, 10:85–97, 1997.

K. Parsons, M. Butavicius, M. Pattinson, A. McCormac, D. Calic, and C. Jerram. Do Users Focus on the Correct Cues to Differentiate Between Phishing and Genuine Emails? In *Australasian Conference on Information Systems*, page 10. AISEL, 2015a.

K. Parsons, A. McCormac, M. Pattinson, M. Butavicius, and C. Jerram. The design of phishing studies: Challenges for researchers. *Computers & Security*, 52:194–206, 2015b.

K. Parsons, M. Butavicius, P. Delfabbro, and M. Lillie. Predicting susceptibility to social influence in phishing emails. *International Journal of Human-Computer Studies*, 128:17–26, 2019.

J. Pastor-Galindo, P. Nespoli, F. G. Mármol, and G. M. Pérez. The not yet exploited goldmine of OSINT: opportunities, open challenges and future trends. *IEEE Access*, 8:10289–10292, 2020.

C. S. Reichardt. Criticisms of and an alternative to the shadish, cook, and campbell validity typology. *New Directions for Evaluation*, 2011(130):43–53, 2011.

M. P. Robillard, D. M. Arya, N. A. Ernst, J. L. Guo, M. Lamothe, M. Nassif, N. Novielli, A. Serebrenik, I. Steinmacher, and K.-J. Stol. Communicating study design trade-offs in software engineering. *ACM Trans. Softw. Eng. Methodol.*, 33(5), 2024.

R. W. Rogers. A Protection Motivation Theory of Fear Appeals and Attitude Change1. *The Journal of Psychology*, 91(1):93–114, 1975.

M. Schmitt and I. Flechais. Digital deception: generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review*, 57(12):324, 2024.

S. C. Sethuraman, D. P. V s, T. Reddi, M. S. T. Reddy, and M. K. Khan. A comprehensive examination of email spoofing: Issues and prospects for email security. *Computers & Security*, 137:103600, 2024.

W. R. Shadish, T. D. Cook, and D. T. Campbell. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company, 2002.

T. Sommestad and H. Karlzén. The unpredictability of phishing susceptibility: results from a repeated measures experiment. *Journal of Cybersecurity*, 10(1):tyae021, 2024.

T. Sommestad, H. Karlzén, and J. Hallberg. A Meta-Analysis of Studies on Protection Motivation Theory and Information Security Behaviour. *International Journal of Information Security and Privacy (IJISP)*, 9(1):26–46, 2015.

N. Stembert, A. Padmos, M. S. Bargh, S. Choenni, and F. Jansen. A Study of Preventing Email (Spear) Phishing by Enabling Human Intelligence. In *European Intelligence and Security Informatics Conference*, pages 113–120. IEEE, 2015.

M. Steves, K. Greene, and M. Theofanos. Categorizing human phishing difficulty: a Phish Scale. *Journal of Cybersecurity*, 6(1):9, 2020.

R. L. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

T. N. Tideman. Independence of clones as a criterion for voting rules. *Social Choice and Welfare*, 4(3):185–206, 1987.

University of Chicago. Ethical Guideline for Online Interviews - Virtual Ethnographic Methods — Class Research Portfolio, 2020. URL `https://voices.uchicago.edu/202003sosc20224/2020/06/25/ethical-guidelines-for-online-interviews/`.

R. Valecha, P. Mandaokar, and H. R. Rao. Phishing Email Detection using Persuasion Cues. *IEEE Transactions on Dependable and Secure Computing*, 19(2):747–756, 2022.

A. Van Der Heijden and L. Allodi. Cognitive triaging of phishing attacks. In *USENIX Security Symposium*, pages 1309–1326. USENIX Association, 2019.

J. Wang, Y. Li, and H. Rao. Coping responses in phishing detection: An investigation of antecedents and consequences. *Information Systems Research*, 28(2):378–396, 2017.

E. J. Williams and D. Polage. How persuasive is phishing email? The role of authentic design, influence and current events in email judgements. *Behaviour & Information Technology*, 38(2): 184–197, 2019.

E. J. Williams, J. Hinds, and A. N. Joinson. Exploring susceptibility to phishing in the workplace. *International Journal of Human-Computer Studies*, 120:1–13, 2018.

L. J. Williams and S. E. Anderson. Job Satisfaction and Organizational Commitment as Predictors of Organizational Citizenship and In-Role Behaviors. *Journal of Management*, 17(3):601–617, 1991.

C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén. *Planning*, pages 89–116. Springer, 2012.

M. Wolff and A. Haase. Viewpoint: Dealing with trade-offs in comparative urban studies. *Cities*, 96, 2020.

O. A. Zielinska, A. K. Welk, C. B. Mayhorn, and E. Murphy-Hill. A temporal analysis of persuasion principles in phishing emails. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 60(1):765–769, 2016.

Table VI: Abuse inbox variable descriptions and unique counts. Source: Authors own work.

Variable	Description	Count
id	Unique id for each reporting instance	8369
reporterAddress	Aliased email address of a reporter	1460
toAddress	Recipient address from the reported email	1921
fromAddress	Sender address from the reported email	3178
subject	Subject of the reported email (if available)	3119
body	Body of the reported email	6503
attachmentName	Attachments' names (if any)	687
attachment	SHA256 hash of attachments (if any)	1118
receivedTime	Timestamp of received time of the reported email (if available)	5102
reportedTime	Timestamp of reported time of the reported email	8356

Table VII: Outcome of manual classification of 131 emails. Source: Authors own work.

		Technical	
		low	high
Contextual	low	56	36
	high	21	18

Table VIII: Outcome of manual classification of 89 phishing emails (removing the non-phishing emails from the total sampled 131 emails). Source: Authors own work.

		Technical	
		low	high
Contextual	low	25	29
	high	19	16

A Details on email classification and reporter clustering

To obtain an insight into the overall email characteristics, we sampled 20 reporters from the dataset (detailed in Table VI) and classified the resulting 131 emails reported by them. We manually classified the emails as phishing vs. not phishing (i.e., legitimate/spam), and high contextual and/or high technical believability. Given the experience of the investigators in the context of UNI, classifying phishing and not phishing emails was a straightforward task with the dataset at hand. The classification of believability, however, was carried out deductively by four investigators by means of the a-priori defined criteria of Section 2. To ensure the quality of the analysis, the four investigators coded the emails by discussing each sampled email, identifying points of disagreements and resolving disagreements, and iteratively updating the analysis of emails (McDonald et al., 2019). Given the limited amount of emails to be coded, two co-coding sessions were enough to resolve disagreements and update the criteria definitions. Out of the 131 emails, 42 (32%) were deemed as not phishing and 89 as phishing (68%), and no ambiguous or unknown type of emails were identified. The believability classification results are reported in Table VII for all 131 sampled emails, and in Table VIII for only phishing emails. The abuse inbox contains emails that vary largely in content as well as in contextual and technical believability. The majority of sampled phishing emails present a low believability on either one of the believability dimensions (19 and 29) or on both (25), and emails high on both dimensions are only 16 (≈18%).

Importantly, we observe a high variability of features in the data that affect contextual and technical believability the most (i.e., `fromAddress`, `subject` and `body`). Previous literature on investigating characteristics of phishing attacks reports similar limitations (Ferreira et al., 2015; Steves et al., 2020). When attempting to characterize more advanced features in phishing emails, previous efforts rely on, for example, massive historical data with ground truth (Ho et al., 2019) or simplify the relevant features to satisfactory levels of approximation within the scope of the study (Van Der Heijden and Allodi, 2019).

The outcome of the review suggests that implementing a machine learning approach to identify high contextual and technical believability reports might be inappropriate for our goals. On the one hand, an unsupervised method to detect similar features might identify sufficiently large phishing or spam campaigns, but it will unlikely identify the less frequent sophisticated emails. On the other hand, *manually* labeling the features that determine a high contextual and technical believability to enable a supervised approach for the purpose of our sampling strategy and to answer the research question would be impractical. For instance, building a training set would require thousands of emails to be labeled by experienced agents with contextual knowledge. Given the relatively static structure of the emails and the specific nature of our dataset of emails that come from only one organization (e.g., identifying targetization towards UNI can be done with a regular expression matching the -short- name of UNI), we apply a rule-based classifier, with the exact rules reported in Table IX. We selected email features that showed a lower variability (e.g., mentions of UNI,

Table IX: Classification rules. Source: Authors own work.

Criterion	Type	Class
A regex with UNI and variations	contextual	high
Otherwise	contextual	low
If any, the payload URL contains: a domain of popular URL shorteners: bit.ly, 1drv.ms, is.gd, tinyurl.com, bit.do, cutt.ly, s.id, rebrand.ly, ht.co, clck.ru, bit.do, rplg.co	technical	high
a domain of popular file hosting services: dropbox.com, drive.google.com, docs.google.com, box.com, mega.nz, onedrive.live.com, forms.office.com, icloud.com, nextcloud.com, spideroak.com, idrive.com, pcloud.com, mediafire.com, tresorit.com, egnyte.com, sugarsync.com, storegate.com, opendrive.com, jungledisk.com, carbonite.com, flipdrive.com, filesanywhere.com, elephantdrive.com, adrive.com, clck.ru	technical	high
a homograph attack in the link based on the strings related to UNI: [redacted for submission] with Levenstein distance is either 1 or 2 from the strings in domain and subdomain	technical	high
a domain of popular free web hosting services*: weebly.com, 000webhost.com, 000webhostapp.com, x10Hosting.com, wix.com, ucoz.com	technical	low
Not any of these: zip, rar, 7z, doc, docx, docm, xls, xlsx, ppt, pptx, pdf, jpeg, jpg, png, gif, exe	technical	high
Otherwise	technical	low

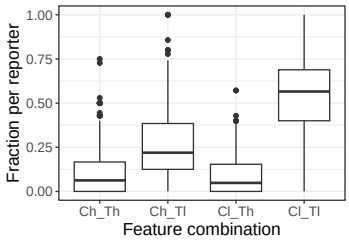
* Overrides previous URL criteria.

Table X: Classification performance for contextual believability. Source: Authors own work.

		Reference	
		low	high
Prediction	low	81	4
	high	11	35

Table XI: Classification performance for technical believability. Source: Authors own work.

		Reference	
		low	high
Prediction	low	70	29
	high	7	25



	Min.	Q1	Mdn	Mean	Q3	Max.
`Ch_Th`	0.000	0.000	0.063	0.107	0.167	0.750
`Cl_Th`	0.000	0.000	0.048	0.088	0.154	0.571
`Ch_Tl`	0.000	0.125	0.219	0.260	0.385	1.000
`Cl_Tl`	0.000	0.400	0.566	0.545	0.689	1.000

Figure 5: Distribution of fractions of reports across the combinations of Contextual and Technical features (Boxplot on the left and detailed statistics on the right).

as opposed to pretexts matching UNI context) and features that are more robust indicators of contextual or technical believability (e.g., free web hosting domains in the URLs signal a low technical believability).

To evaluate the classification performance of the rule-based approach, we used the manually classified emails as the ground truth and ran the naive classifier on it. Tables X and XI report the true positives and negatives, and false positives and negatives for the contextual and technical believability classifiers, respectively. The contextual believability classifier has an accuracy of 88.6% and a precision of 76.1%. The technical believability classifier has an accuracy of 72.5% and a precision of 78.1%.

Fig. 5 shows the distribution of the fractions of reported emails across the combinations of Contextual and Technical features. Figure 6 presents the selection of the optimal number of clusters with the elbow method (Thorndike, 1953) where the within sum of squares was is negligible for $k > 5$.

B Details on participants

Table 7 reports the demographics of the participants.

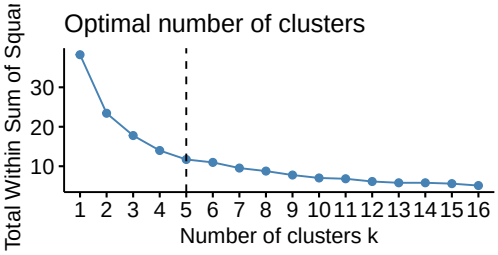


Figure 6: Optimal number of clusters following the elbow method. The plot shows the within the sum of square cost function (Thorndike, 1953) at the varying of the number of clusters k, for an arbitrary max of 16 clusters. The reduction of the within sum of squares was negligible for $k > 5$. Hence, we choose five clusters. Source: Authors own work.

Figure 7: Participants' demographics. Source: Authors own work.

Variable	**Value**	**Freq.**
Gender identity	Female	29
	Male	20
Role	Faculty (Assistant, Associate, Full Professor)	6
	Lecturer	1
	PhD student/PostDoc	3
	Manager	4
	Support staff	27
	Secretary	8
Reporting outside UNI	Yes	25
	No	24
Aware of UNI's ISP	Yes	11
	No	38

C Details on interviews

The interview guide is shown in Listing 1. The guide begins with a brief description of the study goals. We then ask introductory questions to familiarize the interviewee with the conversation format. The study description and introductory questions were crafted to avoid priming the participants on the follow-up questions. Then, Q1 comprises high-level questions aiming at eliciting the rationale and motivations for reporting suspicious emails to the IT inbox. Finally, Q2 requires participants to rank the emails and provide a judgment over the ranking.

We wrapped up the interview by asking about the participants' role at the university from a multiple-choice list derived from the UNI organization chart (cf. Table 7), and by encouraging them to share any comments, feedback, or concerns about the interview and the study.

Listing 1: Interview guide.

```
Consent being recorded
Remind they can pause or stop the interview whenever they want.

The goal of our research at the [research] group of [department] is to explore
what and why our colleagues at [UNI] (and organizations in general) report
suspicious messages to enable the development of better tools and methods against
such threats in the future.

0.1) Have you also reported suspicious emails beyond [UNI]? - e.g., bank,
your email provider etc.
0.2) How do you consider the role of the employee in the protection of
the organization?
0.3) Are you aware of the Information Security Policies at the [UNI]? To what
extent do you believe they are relevant to your security, as opposed to
the university's?

Main questions:

Q1:
1.1) Why would you report a suspicious email?
1.2) What would you say are your main motivations that drive you to report an email
as phishing?

Q2:
2.1) Please sort these four phishing emails in order from the most likely you would
report to IT to the least likely. [participants interacting with Outlook]
2.2) Please explain your ranking
2.3) What characteristics of the message make you more likely to report it?
And less likely?

Closing questions:

These were my questions. Is there any comment or feedback that you would like to
share about this interview or phishing/reporting in general?
```

D Codebook and ranking

The codebook contains a short description of associated codes that (will) form a theme, an inclusion/exclusion criterion, and several (counter) examples. Initial themes were formed upon new cards defining new groups, and with each newly formed theme, a re-sorting of cards was applied to meaningfully assess if and which groups need to be reformed. Codebook definitions were added and/or updated for the new themes (akin to the *constant comparison* (Hoda, 2022)). The codebooks with examples for Q1 and Q2 are reported in Table XII and Table XIII. The unique rankings of emails of Figure 4 is shown in Table XIV.

Table XII: Codebook Q1 with examples. Source: Authors own work.

Theme	Description	Example
`Protect me`	Protect oneself from being a victim, protect own data, computer, or system. This includes avoiding negative consequences, such as data theft or not being able to work.	I want to feel safe in my email [inbox]. (P27) I want to be in the picture. I report it because I want to know if I should worry that my laptop has a virus or what is it (P4)
`Protect others`	Includes the concept of protecting others, people, colleagues, or community as a reason to report a suspicious email. This extends to protecting others' systems and data.	To protect the community (P13) Because it's not only my inbox to receive these emails but also other people (P13) Warning people for certain accounts [those who received it as well] (P12)
`Protect UNI`	Includes the concept of protecting UNI as a reason to report a suspicious email. This extends to protecting systems, networks, and data.	Keep the [UNI] network safe (P14)
`Loyalty`	Explicit expression of loyalty to UNI (as a reason to report).	Loyalty to [UNI] and the community. (P29)
`Help UNI/IT`	Help/assist UNI and/or specifically its IT department in their effort to handle the issue of phishing emails/protecting infrastructure/employees.	Sometimes I know it's only spam, which I also send to abuse, so they block them (P27) Warn [the IT department], so they block access to a website and prevent further damage (P37)
`Help others`	Help/assist colleagues in their efforts to avoid falling for phishing.	Because I can help others with it. (P29)
`Sense of responsibility`	The person feels/knows it's their duty/responsibility to report (or to help/protect UNI/IT/colleagues/data), either because reporting is the norm, a civic duty or simply "the right thing to do".	Then it should be a normal thing to report. [...] I think this is good practice, and should be general practice. (P2)
`Awareness & Experience`	The person is aware of the risks and consequences of phishing/not reporting. This includes previous experience w.r.t. personal experience and/or others' or in the news.	I know the danger phishing can cause, e.g., in Maastricht. (P30) It's one of the biggest risks and nuisances nowadays (P30)
`Ask for confirmation`	Enquiring IT/other colleagues to confirm or refute that an email is phishing as a reason to report.	Also, to ask for confirmation that it is phishing. (P12)
`Efficacy to report`	Statements on (having) the efficacy to report (e.g., because it's easy). Includes the case of knowing how to report and knowing that they should do this (e.g. being told once).	Because I know that there is the abuse@UNI.edu. (P15) Once I was asked to do it, so I do it. (P27)
`Annoyance`	Statements about the feelings of being annoyed or angry at the attacker/attack. Sometimes refers to being annoyed by spam.	Hopefully, I will not get them back in my mailbox because it's so annoying. (P43)
`Fight hackers`	Statements on the desire to fight or punish the attackers, or a feeling of disdain towards the perpetrator as a reason to report.	I don't approve phishing attacks. I want to correct them; (P28)
`Detection`	When they answer the question "how do you detect" instead of "why do you report". For example, reporting because the pretext is strange/suspicious, they do not trust it or the sender is unknown/impersonated.	When I don't believe it, it's too good to be true. (P27) I don't know sender name (P18)

Table XIII: Codebook Q2 with examples. Source: Authors own work.

Theme	Sub-theme	Description	Example
Relevance cues		Mentions of cues or features related to the relevance dimension of the email features or contents.	
	Mismatch between sender, pretext or links.	An inconsistency between email features is perceived.	I look at the sender address. If it's not UNI, I ask IT. (P23) [B_Ch_T1] our IT helpdesk doesn't have this address. Also, I never get emails from SYS ADMN. (P29)
	Premise alignment	The email's content aligns with the target's experiences and expectations.	[C_Cl_Th] [the pretext] "documents for you reference" would seem to me like a reference letter because of my job. (P7) I'd report [D_Ch_Th] because UNI is mentioned there. (P4)
	Targetization	The evaluation of emails' targetization of organization or related entities.	[...] because they want my [Microsoft] account (P12)
Sender cues		Cues related to the sender (name, address) or features related to the sender.	
	Sender untrustworthy	The sender appears either unknown or untrustworthy.	[A_Cl_T1] sender address has .mx (P18) [B_Ch_T1] very recognizable. The email address is suspicious, also the "server@UNI.nl" doesn't exist. (P39)
	Sender disguise as trustworthy or legitimate	The sender features impersonate or disguise as trustworthy.	[A_Cl_T1] looks it's coming from Microsoft, a more trustworthy source. (P33) [C_Cl_Th] comes before the last because it has a more human name, but the link doesn't say anything to me (P1)
Pretext danger perceptions		Mentions of danger perceptions of the pretext (the semantics of the email features and contents).	[D_Ch_Th] much more friendly. (P27) [C_Cl_Th] is tricky. It looks like it's a very important document [. . .] (P39) [B_Ch_T1] because the email can have important information (more important than the less likely emails) (P28)
Persuasion cues		The email pretext employs persuasion techniques, such as urgency or fear.	[A_Cl_T1] because of the sentence "inability to complete the information will render your account inactive." + it's a bit threatening. (P27) It doesn't say to be from UNI, but still gives urgency. (P26)
Content cues		Mentions of email cues or features that relate to the content of the email body.	
	Content has links	The email body has a link or pretext that nudges to follow the link (e.g., click here).	[. . .] it seems like the link/document contains malware (P36). [less] [D_Ch_Th] the sender address is not UNI so report most likely, it also has links in it (P23)
Email etiquette		Mentions of emails features related to the overall look&feel (e.g., tone, syntax, spacing, font, etc.).	[B_Ch_T1] Sender name [has typo] (P10) They are all very bad English, I don't see a difference in that. (P3)
Obvious phishing		Not reporting because the email is considered obvious phishing, SPAM or unwanted, with no further distinction.	[C_Cl_Th] pretty obvious that it is SPAM for everyone (P33) [A_Cl_T1] I'd delete directly, even open it or whatever because it looks like any other email about seminars etc. Wouldn't report. (P11)
Others might fall for it		Reporting because the co-workers might be deceived by the email.	Also, this email has UNI.nl/ICT, but we don't have ICT, but for someone that doesn't know would be tricky (P2)
Report all		Reporting all email without specific reasons (applies to all four emails).	Report them all, since it's easy to report (P29)

Table XIV: Unique email rankings by participants per cluster.

Rank	Tot	clst1	clst2	clst3	clst4	clst5
D_Ch_Th>B_Ch_Tl>A_Cl_Tl>C_Cl_Th	10	2	4	2	1	1
D_Ch_Th>B_Ch_Tl>C_Cl_Th>A_Cl_Tl	6	1	3	1	1	0
D_Ch_Th>A_Cl_Tl>B_Ch_Tl>C_Cl_Th	6	2	1	2	1	0
A_Cl_Tl=B_Ch_Tl=C_Cl_Th=D_Ch_Th	4	0	1	0	2	1
A_Cl_Tl=B_Ch_Tl=D_Ch_Th>C_Cl_Th	3	1	0	1	0	1
B_Ch_Tl>D_Ch_Th>A_Cl_Tl>C_Cl_Th	3	0	0	1	0	2
B_Ch_Tl>A_Cl_Tl>C_Cl_Th>D_Ch_Th	2	1	1	0	0	0
A_Cl_Tl>D_Ch_Th>B_Ch_Tl>C_Cl_Th	2	0	0	0	1	1
D_Ch_Th>A_Cl_Tl>C_Cl_Th>B_Ch_Tl	2	0	0	0	2	0
A_Cl_Tl>B_Ch_Tl>C_Cl_Th>D_Ch_Th	1	1	0	0	0	0
B_Ch_Tl>D_Ch_Th>A_Cl_Tl=C_Cl_Th	1	1	0	0	0	0
C_Cl_Th>D_Ch_Th>A_Cl_Tl>B_Ch_Tl	1	1	0	0	0	0
C_Cl_Th>A_Cl_Tl>B_Ch_Tl>D_Ch_Th	1	1	0	0	0	0
D_Ch_Th>A_Cl_Tl=B_Ch_Tl=C_Cl_Th	1	1	0	0	0	0
A_Cl_Tl>B_Ch_Tl=C_Cl_Th=D_Ch_Th	1	0	0	1	0	0
A_Cl_Tl>C_Cl_Th>D_Ch_Th>B_Ch_Tl	1	0	0	0	0	1
D_Ch_Th>A_Cl_Tl=C_Cl_Th>B_Ch_Tl	1	0	0	0	1	0
D_Ch_Th>B_Ch_Tl=A_Cl_Tl=C_Cl_Th	1	0	0	0	1	0
C_Cl_Th>D_Ch_Th>B_Ch_Tl>A_Cl_Tl	1	0	0	0	0	1
A_Cl_Tl>B_Ch_Tl>D_Ch_Th>C_Cl_Th	1	0	0	0	0	1