safety

MDPI

Review

Cybersecurity Warnings as Safety-Relevant Learning Mechanisms: A Scoping Review of Behavioral Adaptation, Trust Calibration, and Risk Regulation

Eleonora Chiantera, Lorenzo Arciulo and Francesco Di Nocera *

Department of Planning, Design, and Technology of Architecture, Sapienza University of Rome, 00196 Rome, Italy
* Correspondence: francesco.dinocera@uniroma1.it

Abstract

Cybersecurity relies heavily on warning systems to regulate user behavior under uncertainty. These warnings (ranging from browser security dialogs to phishing alerts and enterprise security notifications) do more than convey information: they may alter the conditions under which users select, avoid, verify, report, or override security-related actions. When combined with feedback, they may also contribute to calibrated reliance and safer behavior over time. However, existing research remains fragmented across usable security, human–computer interaction, and safety-related decision-making, and is largely focused on short-term outcomes. As a result, limited attention has been given to how cybersecurity warnings function as risk-control and learning mechanisms within safety-relevant socio-technical systems. This scoping review maps how empirical studies have examined cybersecurity warning systems in relation to behavioral adaptation, trust calibration, and risk regulation, and whether they assess persistence, transfer, or learning over time, identifying recurring design patterns, critical trade-offs, and structural gaps. Following PRISMA-ScR 2018 guidelines, we searched major multidisciplinary and domain-specific databases, with no time frame limits, for empirical studies that examined cybersecurity warnings in relation to learning-relevant behavioral, cognitive, or performance outcomes. Seventeen studies met the inclusion criteria; this number reflects the review's focused conceptual scope rather than the size of the cybersecurity-warning literature as a whole. Eligible studies included experimental, quasi-experimental, field, and mixed-method designs, but no included study assessed persistence or transfer over time. Data extraction focused on warning characteristics, learning and trust mechanisms, user behavior, and security outcomes. Across the included studies, research primarily evaluates immediate responses, such as clicks, choices, response time, and classification accuracy, whereas comprehension and corrective feedback are infrequently assessed, trust calibration is never formally measured, and persistence or transfer over time is assessed in none of the included studies. On this basis, the review proposes a learning-oriented framework for evaluating cybersecurity warnings beyond short-term compliance, emphasizing feedback, calibrated reliance, risk-regulation responses, and direct tests of maintenance and transfer over time.

Keywords: cybersecurity warnings; security alerts; behavioral adaptation; trust calibration; risk regulation; warning design; usable security; human factors; socio-technical systems; safety-relevant learning

Academic Editor: Raphael Grzebieta

Received: 26 May 2026
Revised: 20 June 2026
Accepted: 25 June 2026
Published: 28 June 2026

1. Introduction

Cybersecurity depends not only on the robustness of technical infrastructure or the effectiveness of automated detection systems, but also on the ability of users and operators to interpret risk signals and act safely under conditions of uncertainty. Cybersecurity concerns not only the integrity of networks but also the observable actions of users and operators, such as selecting, avoiding, verifying, reporting, delaying, or overriding security-related actions under uncertainty [1]. This centrality of the human factor emerges in many cybersecurity contexts, especially when the system cannot directly impose a single response and instead requires the user to evaluate a risk signal. Phishing is a particularly clear example of this dynamic, as it represents a primary channel of compromise because it requires the user to decide whether or not to trust content that the system may flag as suspicious but cannot always definitively block [2].

Such decisions are often made under time pressure, with attention divided among multiple tasks. Chowdhury et al. [3] show that time pressure can lead to less deliberate responses, with potentially negative effects on safety. Cybersecurity knowledge should therefore be treated as a necessary but insufficient condition for secure behavior. It can help identify vulnerability profiles, but secure action depends on the task context, the user's prior history, the available warning cues, response costs, feedback conditions, and the consequences arranged for secure and insecure responses [4]. This point emerges particularly clearly in experimental studies that have used phishing detection as a test context to investigate the interaction between users and automated decision-support systems: an explicit description of the system's reliability is not necessarily sufficient to guide behavior, whereas experience and feedback have a more direct impact on the calibration of user trust and response [5,6]. From a safety science perspective, cybersecurity warnings are risk communication and control devices embedded in human–machine interaction. They are deployed in organizations and critical services where failures can escalate into operational disruption, financial losses, and compromised confidentiality. This makes warning design and evaluation relevant not only for usable security, but also for safety management in socio-technical systems. Similarly, when warnings do not align with users' perception of risk or the logic of the task, users could deliberately ignore them, believing that they do not truly apply to the specific situation [7,8]. From this perspective, security knowledge alone is not sufficient to explain behavior in risky situations, neither for ordinary users nor for expert operators. For non-experts, safety decisions depend not only on what they know, but also on how risk is interpreted [9], how capable they feel to act [10], and how safety cues fit the task at hand [7].

Consistently, current warning design recommendations do not seem to fully address why non-experts choose not to protect themselves [10]. The most recent literature considers security knowledge a useful indicator of vulnerability, capable of distinguishing different risk profiles, but not an exhaustive predictor of safe behavior. Tempestini et al. [11] and Di Nocera et al. [4] suggest that knowledge of cyber-threats is relevant for understanding differences among users, but this does not automatically translate into the ability to adopt appropriate responses in real-world situations. These results are consistent with the findings of Kelley and Bertenthal [12], who found that security knowledge alone does not reliably explain users' actual behavior, which is instead influenced by attention to security indicators, familiarity with the site, and task-induced psychosocial stress. For example, a user may know that suspicious links should not be opened, but still click when the warning is generic, interrupts an urgent task, or does not explain which element of the message makes the action risky. Even in cyber defense contexts, however, managing uncertainty remains a structural part of human work: defenders must monitor numerous alerts, identify potential intrusions, and make decisions under time pressure, sometimes

interrupting legitimate activities to protect the network [13]. Consistently, Gay et al. [14] show that, in human–machine teaming contexts, the quality of the response depends on the combination of warning signals, the presence or absence of an attack, and the operational context. From this perspective, cybersecurity appears as a problem of human decision-making, spanning both the experience of the ordinary user and specialized defense contexts. In these processes, uncertainty becomes visible within interfaces through risk signals, such as warnings, alerts, and visual security indicators, which accompany and guide users' decisions in real time. Cues do not eliminate uncertainty, but they can make it manageable, guiding the interpretation of the situation and influencing both the speed and quality of the response [2,8,15]. Decision effectiveness therefore depends not only on the accuracy of the available information, but also on how that information is presented, perceived, and integrated during the task.

Security warnings are discrete notifications designed to alert users to a threat and prompt protective action; compared with other security messages, they are distinguished by their immediacy and strong relevance to the task at hand [16]. This makes warnings a privileged case for usable security research, because they expose the point at which interface design, risk communication, and user behavior converge. Their function, however, should not be limited to conveying information about risk; rather, they operate as antecedent conditions that may guide the user's action by making some responses, such as avoiding, verifying, reporting, or proceeding, more likely in a specific context. In this sense, they operate as risk-regulation interfaces through which the technical system makes a threat visible while leaving the final decision to the user. From this perspective, warnings should not be conceived simply as nudges capable of guiding immediate behavior, but as potential support for the gradual acquisition and maintenance of safe response strategies [17]. Although warnings intervene at the moment of choice, durable behavior change is unlikely to depend on antecedent signaling alone. Durable and flexible behavior change is more likely when users contact programmed consequences for their actions, for example through corrective feedback, confirmation of threat validity, or information about the adequacy of their response [18]. From a behavior-analytic perspective, warnings are better understood as antecedent stimuli or rule-like prompts that specify or signal possible contingencies. They support the selection and consolidation of safer responses only when they are embedded in contingencies that include consequences, such as feedback about response accuracy or threat validity. Cybersecurity warnings do not always operate in life-critical environments, but they mediate decisions in which inappropriate responses may expose individuals or organizations to significant consequences, including loss of valuable assets, unauthorized system access, loss of confidentiality, fraud, identity theft, stolen credentials, phishing, and eavesdropping [9,19]. Their safety relevance lies at the boundary between technical risk detection and human action. Warnings communicate risk and indicate how to avoid hazards [9]. They require users to decide whether to adhere or to proceed despite the risk [8], and they may fail when they do not match users' perceived threat model or task context [7]. Cybersecurity warnings can therefore be designed and evaluated as components of safety-relevant learning environments. Through repeated exposure, explanation, feedback, and design variation, they may help establish more accurate discrimination of threat cues, more appropriate reliance on alerts, and safer responses that persist beyond the prompt itself. Warning systems, therefore, do not merely influence the immediate response to an alert; they operate across at least three interrelated dimensions. First, they may occasion behavioral adaptation, understood here as the observable adjustment of users' or operators' behavior in response to a warning at the moment of decision. This includes actions such as clicking, avoiding, classifying, reporting, overriding, or modifying a security-related behavior. Although repeated exposure may also produce learning or

habituation over time, the present review primarily uses behavioral adaptation to refer to short-term, warning-occasioned behavioral adjustment. Second, they may contribute to calibrated reliance when users' responses vary systematically with the reliability of the signal and of the system that produces it. Experience, feedback, false positives, and false negatives can foster both proportionate trust and forms of distrust or overreliance [5,6,14]. Third, warnings play a role in risk regulation within socio-technical systems, because they do not merely signal danger but intervene at the moment the technical system makes a threat visible and guides the user's decision, altering response costs, available cues, and the probability of observable protective actions, such as verifying, blocking, reporting, delaying, or choosing a safer alternative [2,8,9]. This distinction also requires clarifying the respective meanings of compliance and learning. Compliance refers to the emission of the response recommended by the warning in the current interaction. Learning, by contrast, refers to a change in the probability, accuracy, or flexibility of future responses when similar or partially novel risk conditions are encountered, especially when such change is observed after repeated exposure, after feedback, or when the original warning is absent, less salient, or presented in a different context. For example, a user who closes a phishing email after a warning has complied; a user who later detects a structurally similar phishing attempt without relying on the same warning has shown evidence of learning or transfer. Warnings are therefore not limited to signaling immediate danger. They can also contribute to the progressive development of the repertoires through which users interpret and cope with uncertainty.

This scoping review aims to map how the empirical literature has studied warning systems, with a particular focus on the three dimensions described. The objective is to identify which contexts, methods, outcomes, and design strategies have been most extensively investigated and where any limitations may lie. The review therefore aims to clarify whether and how the empirical literature has studied the relationship between warning systems and behavioral adaptation, how they model trust calibration with respect to the signal and the system that produces it, and how they contribute to risk regulation in different usage contexts. On this basis, it aims to synthesize the available evidence and identify design trade-offs, structural gaps, and lines of development useful for a more integrated conceptual framework for evaluating warning systems as potential components of safety-relevant learning environments. This focus is consistent with a broader critique of usable security research, which has often remained centered on interface comparison and short-term usability outcomes, while paying limited attention to the behavioral processes through which users acquire, maintain, or abandon secure practices [20]. Existing reviews have mapped adjacent areas, including usable security, cybersecurity user experience, and cybersecurity nudging [18,20–22]. However, they do not specifically examine cybersecurity warnings as potential mechanisms through which users learn to interpret risk signals, calibrate trust in alerts, and regulate behavior under uncertainty. The present review contributes to safety science by treating cybersecurity warnings as socio-technical risk-control mechanisms rather than as isolated interface features. From this perspective, warnings are comparable to other safety-relevant signals used in complex systems: they make risk visible, guide action under uncertainty, and influence how users calibrate trust in automated or semi-automated decision-support systems. This framing is particularly relevant for digital environments in which unsafe responses may produce organizational disruption, unauthorized access, loss of confidentiality, or cascading operational consequences. It also connects to long-standing human-factors work on alerting and automation, in which operators' responses to signals of varying reliability are analyzed through signal detection theory (SDT) and through models of appropriate reliance on automation [23–26]. By examining cybersecurity warnings in terms of behavioral adaptation, trust calibration, and risk regu-

lation, the review extends warning research beyond short-term compliance and situates cybersecurity within the broader study of human decision-making, risk communication, and safety management in socio-technical systems.

Within this framework, a cybersecurity warning is not treated as a learning mechanism by itself. It is treated as an antecedent or rule-like stimulus that may occasion a class of security-related responses. Evidence of learning requires changes in later responding, especially when feedback has been provided, when the warning is absent or less salient, or when similar risks appear in new contexts.

2. Materials and Methods

This scoping review was conducted in accordance with the PRISMA-ScR guidelines [27]. The complete PRISMA-ScR 2018 Checklist is provided in Supplementary Materials (Table S1).

2.1. Eligibility Criteria

We focused on peer-reviewed empirical studies that use warnings as a recognizable element of the interfaces, interventions or experimental procedures in digital security contexts and that include an explicit behavioral emphasis on user actions or decisions.

2.1.1. Inclusion Criteria

1. Security domain: The article must explicitly address cybersecurity, such as in the context of phishing, malware, password security, authentication, safe browsing, fraudulent emails, social engineering, data protection, or other cyber risks.
2. Inclusion of an explicit warning or alert: The study must include the presentation of an explicit warning, alert, cautionary message, or risk signal directed at the user.
3. Empirical evaluation: The study had to report at least one behavioral, cognitive, or performance outcome. Behavioral outcomes included clicks, avoidance, reporting, choices, adoption of protective actions, reaction time, or accuracy. Cognitive outcomes included comprehension, risk perception, attention, recall, or awareness, but were coded as learning-relevant only when linked to action or response change.
4. User involvement: The study must involve human users interacting with warnings or alerts or evaluate their responses to such signals in a context of learning, training, instruction, or use.

2.1.2. Exclusion Criteria

1. Lack of relevance in the cybersecurity domain: Studies related to safety in digital environments in a general sense, but not specifically to cybersecurity or IT security. For example, studies on online well-being, content moderation, perceived privacy or general online safety not related to explicit cyber threats will be excluded.
2. Absence of explicit warning: Studies that address cybersecurity without including an explicit warning or alert directed at the user.
3. Purely technical studies: Studies that focus exclusively on technical aspects, such as detection algorithms, automatic classification, filtering systems, or security architectures, without direct user involvement and without evaluating how users interact with warnings or alerts.
4. Lack of empirical data and lack of implementation or evaluation: Theoretical or conceptual articles, opinion papers, reviews, position papers, frameworks, or guidelines that do not present original data; or studies that describe warnings, systems, or interventions solely at the design or conceptual level, without concrete implementation, user presentation, or empirical evaluation.

2.2. Information Sources, Research Strategy and Selection Process

The search was conducted on 10 April 2026 in Scopus, APA PsycInfo, IEEE Xplore and the ACM Digital Library using the query reported in Table 1. The wildcards were used to capture different forms of terms. The query was translated and applied consistently across databases. No time frame limits or additional semantic filters were applied, in order to avoid an overly restrictive selection. Records from IEEE Xplore, the ACM Digital Library, and APA PsycInfo showed substantial overlap with Scopus. After merging all database exports and removing duplicates, 21 records were unique to the non-Scopus databases.

The search strategy deliberately included terms related to learning, training, instruction, and education. This decision constitutes the central eligibility criterion for the review. Since our question concerns whether and how warnings have been studied as mechanisms supporting behavioral adaptation, trust calibration, and risk regulation, studies that evaluate alerts purely as interface artifacts, without any behavioral, cognitive, or performance outcomes related to user response, fall outside this scope. A broader search that omitted terms related to learning, training, and education would have mapped a different and largely already examined literature on the usability of alerts and visual design [18,20–22], which we did not intend to replicate.

The objective of this review is instead to identify empirical studies in which warning systems could be examined as mechanisms supporting behavioral adaptation, trust calibration, and risk regulation. In line with the framing of warnings as potential components of safety-relevant learning environments, the search prioritized studies in which alerts were associated with behavioral, cognitive, or performance outcomes, rather than studies that treated alerts solely as static interface elements. This boundary enhances the conceptual coherence of the review but also introduces a potential limitation: relevant studies on the usability of alerts or visual design may have been excluded if they did not use terminology related to learning, training, instruction, or behavioral change.

The resulting corpus of studies should therefore be read as a focused map of a conceptually delimited body of work rather than as an exhaustive survey of research on cybersecurity alerts. All screening counts reported in Figure 1 refer to the records retrieved through the final search strategy. From the 341 unique records identified, 58 were excluded because they did not meet the peer-review criterion: 7 book chapters, 7 doctoral dissertations, 1 letter, 1 note, 1 short survey, and 41 conference reviews. This resulted in 283 eligible peer-reviewed records to be screened for eligibility, including 134 journal articles, 142 conference papers, and 7 reviews.

During title and abstract screening, records were excluded according to a hierarchical coding procedure. When a record met more than one exclusion criterion, it was assigned to the first applicable category in the following order: lack of relevance to the cybersecurity domain, absence of an explicit warning or alert directed at the user, purely technical focus, lack of empirical data, implementation or evaluation. Potentially eligible records were then examined in full text to confirm the presence of an implemented user-directed warning or alert, empirical user involvement, and extractable behavioral, cognitive, or performance outcomes. Based on this procedure, 91 records were excluded because they addressed security domains other than cybersecurity, such as construction security, cyberbullying, or environmental security. A further 125 records were excluded because they did not include an explicit warning or alert directed at the user. A total of 48 records were excluded because they focused exclusively on technical aspects, such as detection algorithms, automatic classification, or filtering systems. Two articles were excluded because they did not meet the inclusion criterion of user involvement and presence of explicit warning directed at the user. Inter-rater reliability for article inclusion/exclusion decisions was assessed using Cohen's kappa. Agreement between reviewers was substantial to almost perfect: Reviewer

1 vs. Reviewer 2, κ = 0.77; Reviewer 1 vs. final decision, κ = 0.88; Reviewer 2 vs. final decision, κ = 0.89. Discrepancies in inclusion/exclusion decisions were resolved through discussion and consensus among the reviewers.

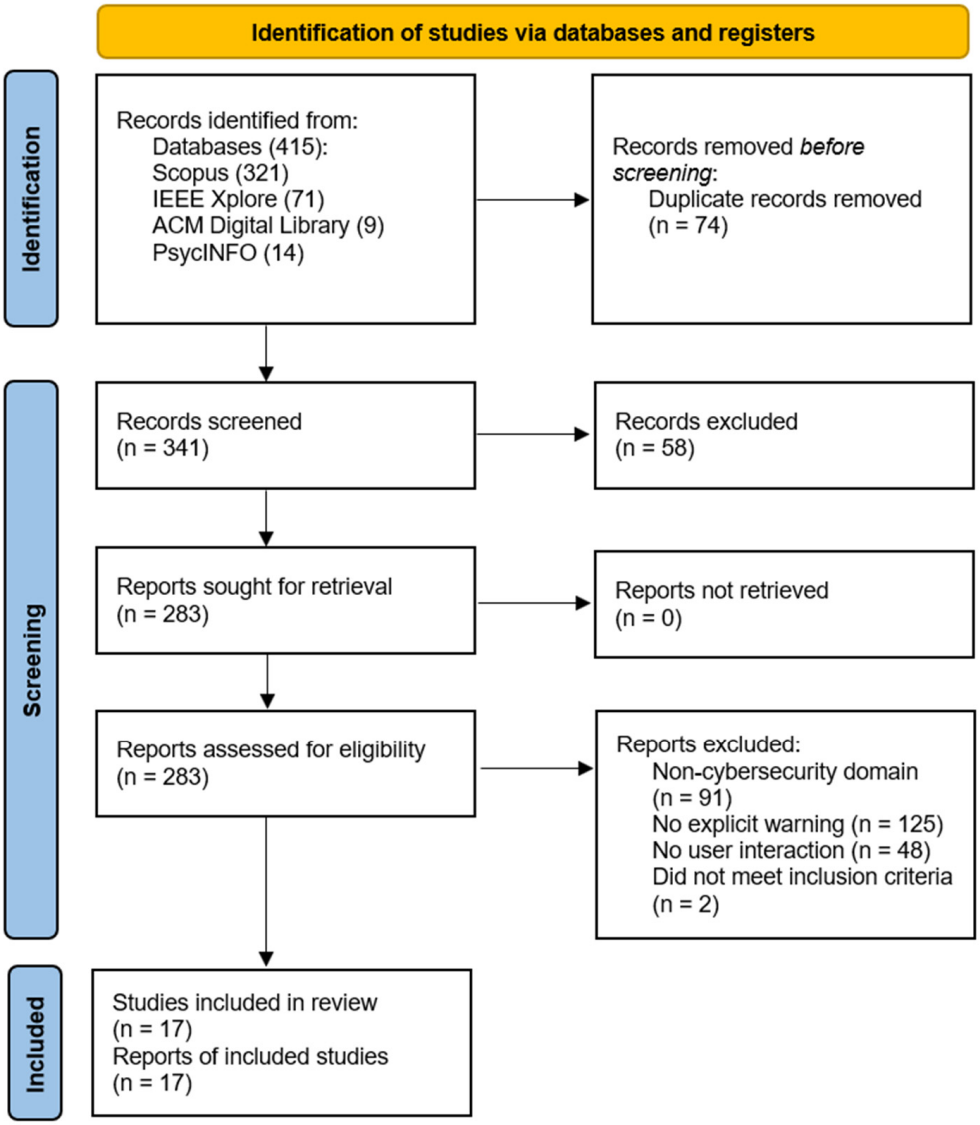

Figure 1. PRISMA flowchart for identification, screening, eligibility, and inclusion.

Table 1. Search queries. *Note. * indicates truncation (wildcard) to retrieve all terms with the same word stem.*

Database	Search Query	Date of Last Search	Records Identified
Scopus	(ABS ((cyber * OR digital)) AND ABS ((securit * OR safety OR attack * OR incident * OR breach * OR threat *)) AND ABS ((warning * OR alert *)) AND ABS ((experiment * OR evaluation * OR test * OR survey *)) AND ABS ((train * OR learn * OR instr * OR educat *)) AND ((user * OR subject * OR participant *)))	10 April 2026	321
IEEE Xplore	(“Abstract”:cyber * OR “Abstract”:digital) AND (“Abstract”:securit * OR “Abstract”:safety OR “Abstract”:attack * OR “Abstract”:incident * OR “Abstract”:breach * OR “Abstract”:threat *) AND (“Abstract”:warning * OR “Abstract”:alert *) AND (“Abstract”:experiment * OR “Abstract”:evaluation * OR “Abstract”:test * OR “Abstract”:survey *) AND (“Abstract”:train * OR “Abstract”:learn * OR “Abstract”:instr * OR “Abstract”:educat *) AND (“Abstract”:user * OR “Abstract”:subject * OR “Abstract”:participant *)	10 April 2026	71
ACM Digital Library	[[Abstract: cyber *] OR [Abstract: digital]] AND [[Abstract: securit *] OR [Abstract: safety] OR [Abstract: attack *] OR [Abstract: incident *] OR [Abstract: breach *] OR [Abstract: threat *]] AND [[Abstract: warning *] OR [Abstract: alert *]] AND [[Abstract: experiment *] OR [Abstract: evaluation *] OR [Abstract: test *] OR [Abstract: survey *]] AND [[Abstract: train *] OR [Abstract: learn *] OR [Abstract: instr *] OR [Abstract: educat *]] AND [[Abstract: user *] OR [Abstract: subject *] OR [Abstract: participant *]]	10 April 2026	9
PsycINFO	AB (cyber * OR digital) AND AB (securit * OR safety OR attack * OR incident * OR breach * OR threat *) AND AB (warning * OR alert *) AND AB (experiment * OR evaluation * OR test * OR survey *) AND AB (train * OR learn * OR instr * OR educat *) AND AB (user * OR subject * OR participant *)	10 April 2026	14

2.3. Data Charting Process

Data charting was conducted on the final set of included articles using a structured Excel extraction sheet. The data charting form was developed to collect both bibliographic information and variables relevant to the review objectives. Specifically, for each included article, we extracted authors, title, year, type of design, setting, number of participants, participant expertise, warning type, warning timing, explanation, specificity, user comprehension, interruption, threat type, recommended action, warning adherence response, and generalization. In addition, three outcome dimensions were charted: behavioral adaptation, trust calibration, and risk regulation. For each dimension, we recorded whether the outcome was empirically assessed in the study. When an outcome was assessed, we further classified the direction of the reported effect.

To reduce ambiguity in data charting, the three focal dimensions of the review were operationalized before synthesis. Table 2 reports the definitions used to determine whether behavioral adaptation, trust calibration, and risk regulation were assessed in each included study. Because the review focused on warnings as potential learning mechanisms, we also coded whether studies provided evidence of generalization or persistence over time. These categories were not treated as mutually exclusive: a single study could assess more than one dimension when its design and outcome measures allowed it.

The four outcome dimensions (behavioral adaptation, trust calibration, risk regulation, and generalization/persistence) were coded against the operational definitions specified a priori in Table 2. Coding was performed by one reviewer and verified in full by a second reviewer against the same definitions, with all disagreements resolved by consensus.

2.4. Data Items

A structured Excel spreadsheet was used to record bibliographic information, study design, setting, sample size, participant expertise, warning type, warning timing, explanation, specificity, user comprehension, interruption level, threat type, recommended action, warning adherence response, behavioral measures, and generalization. The criteria reported in Table 2 were then applied to determine whether each study assessed behavioral adaptation, trust calibration, risk regulation, and generalization or persistence over time. For each assessed dimension, we recorded whether the evidence was based on statistical inference or on descriptive or qualitative observations and classified the direction of the reported effect as improvement, worsening, no change, variable, or unclear. These directional codes summarize the effect as reported in each primary study; because the studies differ in design, sample, setting, and evidentiary basis, and were not appraised for methodological quality, the codes are not validated or mutually comparable effect estimates and should be interpreted in light of these study features.

2.5. Critical Appraisal of Individual Sources of Evidence

No formal critical appraisal of the included sources of evidence was conducted, as the aim of this scoping review was not to determine the effectiveness of specific interventions or to exclude studies based on methodological quality. Rather, the review sought to map empirical evidence on how cybersecurity warnings are used and studied as potential learning mechanisms, with particular attention to behavioral adaptation, trust calibration, and risk regulation over time. Accordingly, the synthesis prioritized the identification of warning characteristics, outcome domains, and conceptual patterns emerging from the literature.

Table 2. Operational coding framework for outcome dimensions. ***Note.*** *The operational definitions are intentionally conservative. Trust calibration was coded as assessed only when the study design allowed reliance to be evaluated against the actual reliability of the warning (e.g., true- vs. false-alert conditions); risk regulation only when a managed-risk strategy beyond a binary response was observable (verification, seeking information, delaying, blocking, reporting, or choosing a safer alternative). These thresholds prevent the over-coding of subjective trust ratings or generic compliance as evidence of calibration or regulation, keeping each code tied to observable, design-supported evidence.*

	Operational Definition	Eligible Indicators	Non-Eligible Indicators	Coding Decision
Behavioral adaptation	Observable adjustment of users' or operators' behavior in response to a cybersecurity warning or alert at the moment of decision.	Clicking or not clicking a link; proceeding or not proceeding after a warning; avoiding a suspicious action; classifying an email, website, or alert as safe or unsafe; reporting a suspicious message; overriding or complying with a warning; changes in response accuracy, response time, or decision quality when directly linked to warning exposure.	General attitudes toward cybersecurity; perceived usefulness or clarity of the warning without an observed or intended behavioral response; cybersecurity knowledge scores alone; awareness or comprehension measures not linked to action.	Coded as assessed when the study included an explicit behavioral or intended behavioral measure related to the user's response to the warning. The direction of the effect was coded as improvement, worsening, no change, variable, or unclear according to the reported findings.
Trust calibration	Alignment between the user's or operator's reliance on the warning and the actual reliability or validity of the alerting system.	Differential responses to true alerts and false alerts; sensitivity to warning reliability; appropriate reliance on valid alerts; reduced overreliance on false positives; reduced underreliance on true positives; behavior across hit, miss, false alarm, or correct rejection conditions.	General trust ratings not linked to objective alert reliability; perceived credibility of the warning without true/false alert conditions; satisfaction with the system; self-reported confidence without evidence of alignment between response and warning validity.	Coded as assessed only when the study design allowed evaluation of whether responses varied according to the actual reliability or correctness of the warning. If trust was measured only as a subjective attitude, but not as alignment with warning reliability, trust calibration was coded as not assessed.
Risk regulation	Observable strategy through which users or operators manage, reduce, verify, or control cyber risk after receiving a warning.	Verifying a suspicious message, link, website, attachment, or alert; seeking additional information; delaying action; blocking a threat; reporting a suspicious email or incident; choosing a safer alternative; modifying the course of action to reduce exposure to risk.	Simple click/no-click responses when no risk-management strategy can be inferred; generic compliance with a warning without evidence of verification, reporting, blocking, or other regulatory action; perceived risk without observable management of that risk.	Coded as assessed when the study provided evidence that the warning influenced how participants managed the risk situation, beyond a binary response. When the study measured only immediate compliance or avoidance without further indication of risk-management strategy, risk regulation was coded as not assessed.
Generalization or persistence over time	Evidence that the effect of the warning extends beyond the immediate interaction in which it is presented.	Follow-up assessments; repeated exposure designs assessing change over time; transfer to new threats, interfaces, tasks, or contexts; persistence of safer behavior after warning removal; longitudinal measures of habituation, learning, or decay.	Single-session responses; immediate post-warning behavior; one-time judgments; claims of learning based only on immediate compliance.	Coded as assessed when the study evaluated whether behavioral change persisted over time or transferred to different contexts. Otherwise, effects were interpreted as immediate responses to warning exposure.

2.6. Synthesis of Results

Charted data were summarized descriptively and narratively. Study characteristics, warning features, and outcome domains were organized in tables to provide an overview of the included evidence. Warning-related variables were compared across studies to identify recurring patterns in warning design, timing, specificity, explanation, interruption level, and recommended action. Outcomes related to behavioral adaptation, trust calibration, and risk regulation were synthesized narratively by describing whether they were assessed and the direction of the reported effects.

3. Results

Overall, 17 studies met the inclusion criteria and were included in the final synthesis (Tables 3 and 4). Throughout this section, findings are reported descriptively: we summarize what each study measured and the direction of the reported effects, without inferring causal relationships between warning characteristics and outcomes. Together, these descriptive patterns provide an overview of the existing evidence and identify issues relevant to the design and evaluation of cybersecurity warnings. Some included studies did not directly assess the focal dimensions as operationalized in this review, despite meeting the broader eligibility criteria. Rather than excluding these studies post hoc, we retained them because their inclusion serves a specific mapping function: they document the extent to which warning research evaluates outcomes, such as perceived clarity, intended behavior, threat appraisal, or usability, that fall outside the behavioral adaptation, trust calibration, and risk regulation framework proposed here. Their presence in the synthesis is therefore not incidental but constitutive of one of the review's central findings, namely that the empirical literature has not yet systematically studied cybersecurity warnings as potential learning mechanisms, even in studies specifically designed to evaluate user responses to warning systems. A descriptive analysis of the included studies by level of expertise revealed that the majority of studies focused on general users, while a minority targeted cybersecurity professionals [28–30]; only one study involved a mixed sample [31]. The studies are predominantly based on experimental methodologies; even in cases where mixed methods are used, these still include an experimental procedure [28,31–33]. In these studies, responses to warnings are primarily measured through immediate behavioral metrics such as clicks, choices, and response time. Clicks capture interaction with potentially risky content. Choices capture the decision outcome or decision quality. Response time captures how quickly participants react to the warning or implement a protective action.

Among the outcomes considered, behavioral adaptation is the most frequently investigated dimension. It is not measured with the same level of evidence across all studies: in some cases, it is supported by statistical analyses [2,12,16,28,30,31,34], while in others it is inferred exclusively from descriptive or qualitative observations [32,35,36]. The studies report mixed results. In some cases, exposure to the warning is associated with improved behavior [2,28,30,32,35], for example by demonstrating greater accuracy in detecting threats [28,30] or reduced susceptibility to phishing and malicious content [2,32]. In other studies, however, the effects are variable [12,36], non-significant [34], or indicate a deterioration in behavioral response [16,31]. This overview shows that behavioral adaptation is widely considered but does not follow a uniform pattern across the included studies. The following synthesis is descriptive and exploratory. The review does not allow causal conclusions about the superiority of a specific warning format. Across the included studies, no consistent pattern links a single warning characteristic to safer behavior. Instead, outcomes vary by context, task constraints, user expertise, and how warning information is integrated into decision-making. With regard to the type of warning used, behavioral adaptation outcomes appear nonspecific: that is, they do not seem to depend systematically on the

mode of presentation of the warning, but vary even within similar formats. For example, email warnings/banners are associated with positive outcomes in some studies [2,32], but do not in themselves guarantee improvement. Similarly, dashboards/enterprise alerts are associated with behavioral improvements in some cases [28,30], while in others behavioral performance is not directly assessed or does not show changes clearly attributable to the alert format [29,37]. Other modalities, such as browser dialogs [12,31,38], pop-ups [16], embedded/in-context warnings [34,39], or sonifications [33], show mixed and not directly comparable outcomes. Overall, therefore, the type of warning helps describe the format of the intervention, but does not seem to explain the direction of behavioral adaptation on its own. Regarding specificity, there are generic, threat-specific, and context-specific warnings, and in some cases, multiple combinations. Generic warnings signal a risk in general terms, threat-specific warnings identify the type of threat being reported, while context-specific warnings link to concrete features of the situation or the user's action. Generic warnings are associated with both non-significant outcomes [34] and worsening outcomes [16]. Threat-specific warnings appear in studies with positive outcomes [32], but also in studies reporting worsening outcomes [31] or in which behavioral adaptation is not directly assessed [14,33,37]. Context-specific warnings were the category most frequently associated with favorable short-term behavioral outcomes. Three studies reported statistically supported improvements, including reduced click-through rate [2], greater decision-making accuracy [30], and better support for alert triage [28]. A fourth study reported mainly descriptive or qualitative positive evidence [35], while one study using this coding reported variable results [12]. Because contextual specificity was not isolated experimentally across studies, this pattern is interpreted descriptively.

Mixed forms (threat-specific + context-specific) also show ambiguous results, with variable outcomes [36] or outcomes not directly assessed [39]. Consequently, specificity appears useful for describing the informational content of the warning, but the available evidence does not allow for the identification of a consistent association with the direction of behavioral change.

Regarding interruption level, blocking warnings are not consistently associated with improvements in behavioral adaptation. Some blocking warnings are associated with positive outcomes [2,35], whereas others yield worsening outcomes, as in Bostan [31] and Vance et al. [16]. Findings for non-blocking warnings were not uniform, including both improvements [28,30,32] and variable outcomes [12,36]. Overall, the level of interruption describes the extent to which the warning interferes with the user's ongoing action, but it does not, by itself, predict the direction of behavioral change.

The level of explanation also varies considerably across studies and does not show a linear association with the direction of behavioral adaptation. Warnings ranged from absent to minimal, detailed, or multiple explanations; however, outcomes varied substantially within each level. Some positive results emerge in studies with minimal explanations [32,35], while others are reported in the presence of detailed explanations, such as in the XAI alerts (security alerts generated by AI-based threat detection systems) of Maram et al. [28]. Studies that manipulate the level of explanation show that its effect depends on how the information is incorporated into the decision-making process. Greco et al. [2], for example, show that explanations reduce the click-through rate especially when the warning is presented after the email is opened, but they also increase the perceived cognitive load and are less suitable in cases of false positives. In Rasmussen et al. [30], however, justifiable recommendations—that is, those accompanied by a justification—improve analysts' accuracy compared to conditions with limited or absent recommendations. Overall, therefore, the amount of information provided by the warning appears to contribute to the quality

of decision support, even if it does not seem to allow for a stable prediction of behavioral improvement on its own.

Comprehension is assessed directly in only a limited number of studies. In cases where it is measured, findings are uneven: some studies report high or moderate comprehension, often in the presence of explanations or cues designed to clarify the warning's meaning [2,33,38], while others highlight lower levels of comprehension, primarily based on qualitative data collected through think-aloud protocols or interaction observations [35,36]. This difference may also depend on the type of measure used. Self-reported measures, such as judgments of clarity, familiarity, or perceived comprehensibility, may yield a more subjective estimate of comprehension [2,38]. In contrast, think-aloud protocols and interaction observations allow for a more direct assessment of whether the user actually understands the warning, as they document what the user does, says, or misinterprets while using the tool [35,36]. This variability does not correlate linearly with the direction of behavioral adaptation: some studies with moderate or high comprehension report positive outcomes, while others primarily assess perceptions, clarity, or interpretation of the warning without directly measuring behavioral change. In most studies, however, comprehension is not assessed [12,14,16,29–32,34,37,39,40]. This limits the ability to determine to what extent the observed behavioral effects are actually linked to users' understanding of the warning.

The timing of the warning is predominantly during the action, while studies in which the warning is presented before the action or at multiple times are less frequent. Even for this characteristic, no consistent pattern emerges in the direction of behavioral adaptation. Warnings delivered during the action appear in studies reporting positive [28,30,35], variable [12,36], and negative outcomes [16,31]. Warnings presented before the action also show mixed results, with positive outcomes in some cases [32] and non-significant results in others [34]. Among studies with multiple timing conditions, Greco et al. [2] provide a particularly relevant example, as they compare warnings displayed before opening the email and warnings displayed after reading, finding that the latter, especially when accompanied by explanations, reduce risky behavior to a greater extent.

The absence of corrective feedback on risk-related behavior also emerged as a relevant aspect. Across the studies included, there is no clear evidence that participants received explicit feedback on the accuracy of their response, such as whether clicking or avoiding a link, proceeding or interrupting an action, reporting an email, or classifying an alert as genuine or false constituted the correct response to the threat.

Table 3. Descriptive characteristics of the studies included in the review.

Authors	Title	Subjects Expertise	Sample Size	Warning Explanation	Warning Timing	Feedback	Behavioral Measure
Blancaflor et al. [32]	AI-Driven Phishing Detection: Combating Cyber Threats Through Homoglyph Recognition and User Awareness	General users	50; 20	Minimal	Before action	Absent	Click
Bostan [31]	Implicit learning with certificate warning messages on SSL web pages: what are they teaching?	Mixed sample	38; 32; 16	Minimal	During action	Absent	Awareness; Choice
Datta et al. [33]	Warning users about cyber threats through sounds	General users	26	Absent	During action	Absent	Intended behavior Threat appraisal
Desolda et al. [38]	Explanations in warning dialogs to help users defend against phishing attacks	General users	150	Detailed	During action	Absent	Intended behavior
Gay et al. [14]	Operator Suspicion and Human–Machine Team Performance under Mission Scenarios of Unmanned Ground Vehicle Operation	General users	32	Minimal	During action	Absent	Choice, Response Time
Greco et al. [2]	Enhancing Phishing Defenses: The Impact of Timing and Explanations in Warnings for Email Clients	General users	900	Multiple	Multiple	Absent	Click, Choice
Howell et al. [34]	Engaging in cyber hygiene: the role of thoughtful decision-making and informational interventions	General users	164	Minimal	Before action	Absent	Choice
Kelley & Bertenthal [12]	Attention and past behavior, not security knowledge, modulate users' decisions to login to insecure websites	General users	173	Absent	During action	Absent	Choice
Maram et al. [28]	XAI-Driven Security Systems: Enhancing Trust and Clarity in Automated Threat Detection and Response	Security professionals	30	Detailed	During action	Absent	Choice, Response time
McRee [29]	Improved Detection and Response via Optimized Alerts: Usability Study	Security professionals	95	Minimal	During action	Absent	Intended behavior
Rasmussen et al. [30]	Nimble cybersecurity incident management through visualization and defensible recommendations	Security professionals	18	Multiple	During action	Absent	Choice, Response Time

Table 3. *Cont.*

Authors	Title	Subjects Expertise	Sample Size	Warning Explanation	Warning Timing	Feedback	Behavioral Measure
Roden & Layman [37]	Cry wolf: Toward an experimentation platform and dataset for human factors in cyber security analysis	Unclear	51	Minimal	During action	Absent	Choice, Response Time
Sharevski et al. [39]	Phishing with Malicious QR Codes	General users	173	Minimal	During action	Absent	Choice
Shava & Van Greunen [40]	Factors affecting user experience with security features: A case study of an academic institution in Namibia	General users	53	Not specified	Not specified	Absent	Habitual behavior
Stembert et al. [35]	A Study of Preventing Email (Spear) Phishing by Enabling Human Intelligence	General users	24	Multiple	During action	Absent	Click, Choice
Vance et al. [16]	The fog of warnings: how non-security-related notifications diminish the efficacy of security warnings	General users	609; 25; 234	Absent	During action	Absent	Choice, Response time
Zheng & Becker [36]	Checking, nudging or scoring? Evaluating e-mail user security tools	General users	27	Multiple	During action	Absent	Click, Choice

Table 4. Coded outcome dimensions of the studies included in the review. ***Note***. *Each label refers to the effect observed in a specific study and in its particular setting, including whether the study was conducted in a laboratory context or in a more ecological environment. Therefore, the labels should not be interpreted as context-independent properties of the warning itself. In particular, differences in ecological validity mean that the same nominal label may refer to outcomes that are not directly comparable across studies. Information on study design, sample, setting, and participants' expertise is reported in* Table 3.

Authors	Title	Behavioral Adaptation	Trust Calibration	Risk Regulation	Generalization or Persistence over Time
Blancaflor et al. [32]	AI-Driven Phishing Detection: Combating Cyber Threats Through Homoglyph Recognition and User Awareness	Improvement	Not Assessed	Not Assessed	Not Assessed
Bostan [31]	Implicit learning with certificate warning messages on SSL web pages: what are they teaching?	Worsening	Not Assessed	Not Assessed	Not Assessed
Datta et al. [33]	Warning users about cyber threats through sounds	Not Assessed	Not Assessed	Not Assessed	Not Assessed

Table 4. *Cont.*

Authors	Title	Behavioral Adaptation	Trust Calibration	Risk Regulation	Generalization or Persistence over Time
Desolda et al. [38]	Explanations in warning dialogs to help users defend against phishing attacks	Not Assessed	Not Assessed	Not Assessed	Not Assessed
Gay et al. [14]	Operator Suspicion and Human–Machine Team Performance under Mission Scenarios of Unmanned Ground Vehicle Operation	Not Assessed	Variable	Variable	Not Assessed
Greco et al. [2]	Enhancing Phishing Defenses: The Impact of Timing and Explanations in Warnings for Email Clients	Improvement	Variable	Not Assessed	Not Assessed
Howell et al. [34]	Engaging in cyber hygiene: the role of thoughtful decision-making and informational interventions	Not significant	Not Assessed	Not Assessed	Not Assessed
Kelley & Bertenthal [12]	Attention and past behavior, not security knowledge, modulate users' decisions to login to insecure websites	Variable	Not Assessed	Not Assessed	Not Assessed
Maram et al. [28]	XAI-Driven Security Systems: Enhancing Trust and Clarity in Automated Threat Detection and Response	Improvement	Not Assessed	Not Assessed	Not Assessed
McRee [29]	Improved Detection and Response via Optimized Alerts: Usability Study	Not Assessed	Not Assessed	Not Assessed	Not Assessed
Rasmussen et al. [30]	Nimble cybersecurity incident management through visualization and defensible recommendations	Improvement	Not Assessed	Improvement	Not Assessed
Roden & Layman [37]	Cry wolf: Toward an experimentation platform and dataset for human factors in cyber security analysis	Not Assessed	Variable	Not Assessed	Not Assessed
Sharevski et al. [39]	Phishing with Malicious QR Codes	Not Assessed	Not Assessed	Not Assessed	Not Assessed
Shava & Van Greunen [40]	Factors affecting user experience with security features: A case study of an academic institution in Namibia	Not Assessed	Not Assessed	Not Assessed	Not Assessed
Stembert et al. [35]	A Study of Preventing Email (Spear) Phishing by Enabling Human Intelligence	Improvement	Not Assessed	Improvement	Not Assessed
Vance et al. [16]	The fog of warnings: how non-security-related notifications diminish the efficacy of security warnings	Worsening	Not Assessed	Not Assessed	Not Assessed
Zheng & Becker [36]	Checking, nudging or scoring? Evaluating e-mail user security tools	Variable	Not Assessed	Variable	Not Assessed

A further factor concerns the presence of a recommended action, that is, an explicit indication of the protective behavior the warning should promote. Not all studies clearly specify which behavior the warning should promote: in some cases, the recommended action is absent [12,14,29,33,37,39], while in others it is formulated more generally, such as avoiding phishing or adopting cyber hygiene behaviors [32,34]. In other studies, however, the action is more specific, for example, not proceeding to a suspicious site [31,38], not clicking on a phishing link [2], denying or avoiding a risky request [16], or classifying or responding to an alert [28,30], or verifying, blocking, or reporting a suspicious email [35,36]. However, adherence to the recommended action has been studied less frequently than the mere presence of the behavioral instruction. In the few studies where it is evaluated, adherence to the recommended action is sometimes associated with favorable outcomes such as a reduction in risky behaviors or an improved alert response [2,30,32], but the effect may be non-significant or influenced by individual characteristics [34]. In other cases, adherence is variable or negative. In studies with variable outcomes, this inconsistency appears to be linked primarily to the understanding of the proposed tools: for example, reporting and contextual suggestions can support the evaluation of the email, but they are not always understood or applied uniformly [35]. Similarly, check buttons, suspicion scores, and nudges can facilitate verification of a message's legitimacy, but some users ignore them, perceive them as too intrusive, or do not consistently integrate them into their decision-making process [36]. In the case of overrides, however, the negative outcome appears more closely linked to the design characteristics of the warning: when the warning resembles ordinary notifications or requires habitual interactions, the likelihood of it being ignored increases, while more distinctive methods, such as sliders or drag-and-drop, reduce this tendency [16]. Furthermore, in several studies, adherence to the recommended action is not directly assessed, making it difficult to distinguish between exposure to the warning, adherence to the recommendation, and subsequent behavioral change [14,29,31,33,37–39].

Risk regulation and trust calibration, on the other hand, are assessed less frequently. When risk regulation can be reconstructed, it involves changes in how participants manage the risk situation, for example by halting immediate action [35], further verifying the legitimacy of the message [36], or seeking additional information before responding [14]. Warnings are often implemented as decision-making criteria, helping users decide whether to proceed, verify further, block, or report the threat.

Trust calibration emerges even more marginally and is not formally measured through a specific index. In this context, a formal measure of trust calibration would require quantifying whether users' or operators' reliance on the warning varies in accordance with the actual reliability of the alert, for example across true-alert and false-alert conditions. It can only be inferred descriptively or indirectly in studies that include warnings or alerts with varying reliability, such as hit/false alarm or true/false alert conditions. Framing alert responses in terms of hits, misses, false alarms, and correct rejections situates trust calibration within signal detection theory, which distinguishes sensitivity from response bias in reactions to signals of uncertain validity. The automation-trust literature complements this with the constructs of appropriate reliance, overtrust, and distrust [24,26], offering a vocabulary for the misalignments between reliance and reliability that we observe descriptively across studies. In these cases, the outcome always appears variable; users or operators do not respond to the warning in a uniformly calibrated manner: sometimes the response is consistent with the actual presence of the threat, while in other cases responses emerge that are less aligned, excessively cautious, or not sufficiently sensitive to the absence or presence of risk [2,14,37].

The generalizability of the effects is not assessed in the included studies. In no case is it verified whether the behavior observed in response to the warning persists over time

or transfers to different contexts, threats, or interfaces. The available results therefore primarily describe immediate responses to the warning, observed within the specific task or experimental scenario.

Taken together, the mapped evidence converges on three findings. First, behavioral adaptation is widely measured but heterogeneous, and no single warning feature consistently predicts safer behavior. Second, one mechanism that could help transform immediate compliance into learning, corrective feedback linking the user's response to the true state of risk, is essentially absent across studies. Third, trust calibration is never formally measured, and persistence or transfer is assessed in none of the included studies. These findings are organized in a learning-oriented framework (Figure 2) in which the absent measures of feedback, trust calibration, persistence, and transfer are treated as unobserved pathways requiring direct empirical evaluation. In this framework, cybersecurity warnings and their design characteristics, such as timing, specificity, explanation, disruption level, and recommended action, function as risk signals that can guide behavior at the moment of decision.

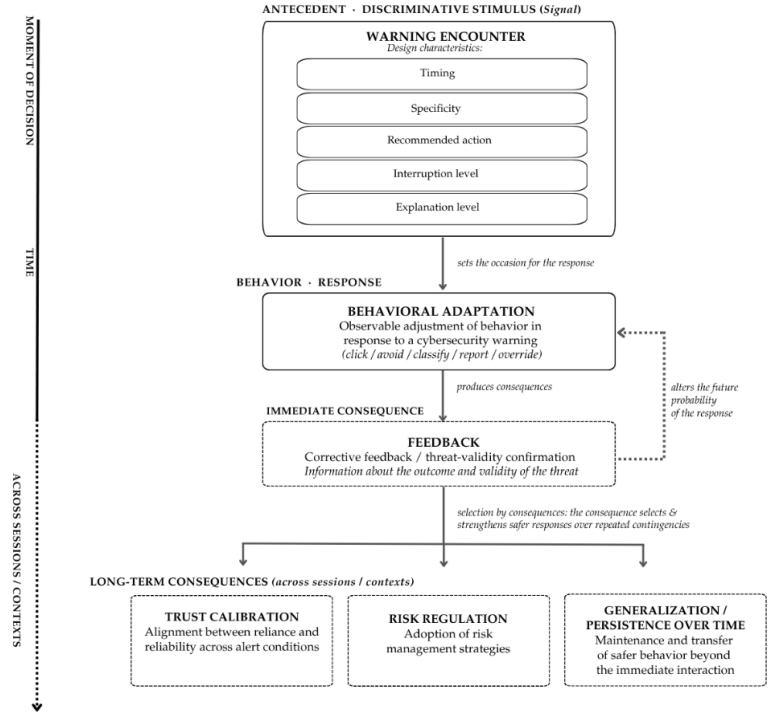

Figure 2. A proposed three-term contingency framework for evaluating cybersecurity warning responses.

Exposure to the alert may produce behavioral adaptation, understood as adjustment of behavior in response to the warning. For this adaptation to become learning, users need feedback about the relation between the warning, their response, and the actual state of risk. Repeated exposure to such contingencies may support trust calibration by helping users align reliance with alert reliability. It may also support risk regulation by strengthening strategies such as verifying information, seeking additional cues, delaying action, reporting, or choosing safer alternatives.

4. Discussion

Behavioral adaptation is the most frequently investigated outcome in the literature on cybersecurity warnings, but the available evidence remains fragmented. This heterogeneity can be organized around three families of candidate moderators that recur across the

included studies: warning design, population, and task context. At the design level, features such as timing, specificity, explanation, interruption, and recommended action shape user response, but none determines its direction on its own. At the population level, samples range from general users to security professionals, and expertise plausibly affects how a warning is interpreted. At the task and context level, time pressure, site familiarity, perceived reputation, and ecological versus laboratory settings may influence whether and how a warning is integrated into the decision process. Because primary studies typically vary on several of these dimensions at once and rarely isolate a single factor, the observed patterns are best understood as multiply determined. Some studies report reductions in risky behavior or improvements in threat detection after exposure to warnings, including reduced phishing susceptibility and improved alert management [2,28,30,32,35]. Other studies report variable responses [12,36], non-significant behavioral effects [34], or negative outcomes, including habituation and warning disregard [16,31]. These mixed findings do not indicate that warnings are ineffective. Rather, they show that warning effectiveness depends on whether the signal comes to occasion appropriate responses within a specific decision context, including avoiding, verifying, reporting, blocking, delaying, or proceeding when the alert is invalid. A warning may support safer behavior when it clarifies the risk, fits the user's task, and offers actionable information. The same warning may fail when it increases processing demands, conflicts with prior expectations, appears too often, or is not perceived as relevant.

This interpretation is consistent with the previous literature showing that cybersecurity decision-making is shaped by psychological and cognitive processes involved in risk evaluation and warning response. Prior work highlights the role of threat perception in shaping how individuals evaluate cyber risks [41], attention allocation in determining whether warning signals are noticed and processed [42], mental models in influencing how non-experts understand cyber threats and protective actions [10], and the integration of explicit and implicit cues during phishing evaluation under uncertainty [15]. A warning response should not be attributed to the warning stimulus alone. It reflects the interaction between the warning, the user's previous experience, the credibility of the source, the demands of the ongoing task, response costs, and the consequences that have followed similar responses in the past.

Warnings could also increase processing demands. Users often need to interrupt their ongoing activity, interpret the signal, infer the type and relevance of the threat, and decide whether a protective response is needed. This is consistent with work suggesting that warning information may overwhelm users [33] and that conflicting risk cues can increase processing demands and slow decision-making during phishing detection [15]. Repeated exposure to similar warnings may further reduce attention and responsiveness, especially when alerts become familiar and trigger automatic responses. Recent studies suggest that repeated exposure to false or excessive cybersecurity alerts may contribute to reduced vigilance, alarm desensitization, and delayed responses [43]. Vance et al. [42] show that attention to warnings declines across repeated exposure and that warning adherence decreases in a longitudinal field experiment, while Amran et al. [19] identify habituation as a central mechanism underlying users' tendency to ignore recurrent security warnings.

A related issue is warning fatigue. Disruptive, frequent, poorly relevant, or non-actionable alerts may become burdensome because they repeatedly require users to stop, shift attention, and evaluate whether the alert is credible and actionable [44]. Users' responses are also shaped by context interpretation, including website familiarity, perceived reputation, prior experience, and the source of the alert [15,45]. Warning effectiveness therefore depends not only on the presence of a risk signal, but also on whether the signal is perceived as relevant, credible, and worth acting on.

Beyond the heterogeneity of behavioral outcomes, the reviewed literature reveals a structural limitation in how cybersecurity warnings are currently conceptualized and studied. Most studies assess immediate behavioral responses to the warning, such as clicks, choices, classification accuracy, or response time. These behaviors are usually observed within a single experimental task or interaction. By contrast, comprehension is assessed less frequently, trust calibration is never formally measured, and persistence or transfer over time is assessed in none of the included studies. This asymmetry suggests that current warning research is better developed for evaluating short-term compliance than for evaluating learning, calibrated reliance, or durable risk-management strategies.

This distinction is important. A warning can produce immediate compliance without producing learning. For example, a highly salient or intrusive warning may reduce click-through rates during an experiment, but the same user may still be unable to identify similar phishing cues when the warning is absent, less salient, or presented in another interface. In that case, the warning has controlled the immediate response, but it has not necessarily helped the user develop a more stable capacity to recognize and manage risk. If warnings are expected to function as safety-relevant learning mechanisms, their evaluation should therefore go beyond whether the user complied in the current interaction. It should also examine whether users understand why an action is risky, whether they can use that knowledge in later situations, and whether their reliance on alerts becomes aligned with the actual reliability of the system.

The near absence of corrective feedback is especially important because it limits both learning and trust calibration. In most included studies, users are exposed to a warning and make a decision, but they are not told whether their response was appropriate given the actual threat. They rarely receive information about whether the warning corresponded to a genuine threat, whether the alert was a false positive, or why a particular response was safer. Without such feedback, users have few opportunities to refine risk evaluation, calibrate trust in the signal, or understand when compliance and non-compliance are warranted. The warning may therefore remain an isolated cue rather than part of an iterative learning process.

This point also clarifies the difference between warning adherence and trust calibration. Trust calibration requires alignment between reliance on the warning and the actual reliability or validity of the alert, not merely perceived credibility, self-reported trust, or immediate compliance. In the included studies, trust calibration is not formally measured through specific indices. It can only be inferred descriptively in studies that include true-alert and false-alert conditions, or hit and false-alarm structures [2,14,37]. In these cases, responses appear variable: users or operators do not always rely on alerts in proportion to their actual validity. This variability is central for safety, because overreliance may lead users to accept false positives or delegate too much judgment to the system, whereas underreliance may lead them to ignore valid alerts. Future warning studies should therefore measure not only perceived credibility or self-reported trust, but the degree to which behavior changes appropriately as alert reliability changes.

The absence of longitudinal or transfer assessment further limits the interpretation of warning effectiveness. The included studies are almost uniformly cross-sectional and single-session. None implement follow-up, repeated-exposure-with-withdrawal, or transfer designs. As a result, they cannot distinguish behavior maintained while the warning is present from safer behavior that persists after the warning is removed. An immediately safe response may depend on the specific features of the warning, the experimental task, or the temporary salience of the alert. It may not indicate stable security competence. If warnings are to be understood not only as reactive alerts but also as potential learning environments, future evaluations should examine whether warning effects persist across

sessions, transfer to different threats or interfaces, and support safer decisions when the warning is less salient or absent.

The present findings therefore support a broader conceptualization of cybersecurity warnings. The relevant question is not only whether a warning prevents a risky action at a specific moment—it is also whether the warning helps users establish response criteria that remain useful when warnings are ambiguous, absent, or only partially informative. From this perspective, warning systems should be designed and evaluated as socio-technical mechanisms that support adaptive risk regulation over time. Their function is not limited to interruption. They can also support interpretation, feedback, trust calibration, and progressively more autonomous decision-making.

Contextual specificity appears particularly relevant in this regard. In the included corpus, context-specific warnings are the category most frequently associated with favorable short-term behavioral outcomes, including reductions in click-through rate in phishing scenarios [2], support for alert triage [28], and improvements in analysts' decision accuracy [30]. Because the available studies do not isolate contextual specificity as a single factor, this pattern should be interpreted as a descriptive tendency rather than as causal evidence. It nevertheless identifies contextual specificity as a design dimension that warrants systematic examination in future experimental work with controlled manipulations. Warnings that connect risk to concrete features of the current action, task, or interface may be more likely to support interpretation than generic warnings that simply announce danger.

Explainable and contextualized warning approaches are consistent with this direction. They were proposed to overcome the limitations of generic warnings and to support the construction of more adequate mental models of cyber threats [46]. A learning-oriented warning system may also require feedback mechanisms that support trust calibration. If users are informed whether a warning reflected a genuine threat, whether their response was appropriate, or whether the alert was a false positive, they may be better able to align future responses with the reliability of the warning system. In this way, feedback can help transform isolated warning encounters into repeated learning opportunities.

Adaptive scaffolding is another promising design dimension. Users differ in cybersecurity knowledge, prior exposure, and ability to interpret threat cues. Non-experts may benefit from more explicit explanations and clearer links between the warning and the risky feature of the situation. More experienced users may benefit from prompts that encourage active evaluation rather than passive compliance. This possibility is consistent with evidence that non-experts differ in how they understand cyber threats and that existing design recommendations often fail to address why they do not protect themselves [10]. It is also consistent with findings suggesting that explainable warnings may be especially useful for users with lower initial cybersecurity proficiency [46]. Adaptive scaffolding therefore emerges as an untested but theoretically relevant design dimension. Future studies should evaluate whether varying warning explicitness according to user knowledge or behavioral history promotes more autonomous threat recognition across sessions and contexts.

4.1. Practical Implications for Warning Design

The patterns identified in this review suggest four candidate priorities for warning design: outcome feedback, contextual specificity, management of false-positive costs, and adaptation to user expertise. However, because the included studies do not directly demonstrate persistence, transfer, or autonomous learning after exposure to warnings, these priorities should not be interpreted as established design guidelines. Rather, they are proposed as testable hypotheses and as criteria for future evaluation. Accordingly, each priority is formulated below as a proposition to be assessed through controlled studies and context-sensitive implementation in local practice.

First, warnings should be coupled with outcome feedback delivered after the user has responded. This means providing information that reveals whether the action taken, for example clicking, avoiding, proceeding, reporting, or verifying, was appropriate given the true state of the threat. This feedback is the mechanism most clearly absent across the included studies and the one most directly tied to learning. By arranging explicit consequences for the user's response, such as information about response accuracy and threat validity, feedback can support the calibration of future responses rather than producing isolated compliance. It is also a precondition for trust calibration, because users can align reliance with the system's actual reliability only if they learn whether an alert corresponded to a genuine threat or to a false positive.

Second, alerts should be tied to concrete features of the user's current action, task, or interface rather than phrased only in generic terms. Context-specific warnings were the configuration most often associated with favorable behavioral outcomes in this corpus. Although this pattern should be tested through controlled manipulations, it suggests that warnings may be more useful when they explain what feature of the current situation is risky and what action should be taken.

Third, warning design should anticipate the costs of false positives. False positives may increase cognitive load, reduce perceived relevance, and erode adherence over repeated exposure. Warning systems should therefore be evaluated not only in terms of how strongly they interrupt risky action, but also in terms of how they affect workload, attention, and future responsiveness when alerts are wrong or only partially relevant.

Fourth, warning explicitness should be adapted to user expertise and behavioral history when possible. Novices may need more explicit explanations, examples, and recommended actions. Experienced users may benefit from more concise warnings that prompt verification and active evaluation. Finally, warnings should be evaluated not only by immediate compliance, but also by comprehension and by whether safer behavior is retained when the warning is less salient or absent. These priorities should be implemented with local evaluation, because their effectiveness is likely to depend on task demands, user expertise, warning reliability, and organizational context.

4.2. Limitations

Several limitations of the present scoping review should be acknowledged. A first limitation concerns the scope of the search and its consequences for the included sample. Because eligibility required warning exposure to be paired with learning-, training-, or instruction-related framing, the review necessarily over-represents studies that explicitly adopt this vocabulary and under-represents studies that examine warnings purely as interface or compliance phenomena. This may include work on warning adherence, habituation, warning fatigue, or visual usability that does not use learning-related terms. The resulting set of seventeen studies should therefore be read as a purposive, conceptually bounded map rather than as a representative cross-section of the entire cybersecurity-warning literature.

This has a specific interpretive implication. Statements about how frequently particular outcomes are measured, especially the absence of formal trust calibration measures and the absence of persistence or transfer assessments, refer to this bounded corpus and should not be generalized to the whole field without further review. At the same time, the modest number of included studies is informative. It shows that cybersecurity warnings are rarely studied through a learning-oriented lens, even when studies evaluate user responses to warning systems. A complementary review adopting broader eligibility criteria would be useful to map warning studies that fall outside the present scope.

A second limitation concerns the heterogeneity of the included studies. The reviewed literature spans different contexts, including phishing detection, browser security warnings,

malware alerts, enterprise security dashboards, and cyber defense systems. Studies also vary in user expertise, warning characteristics, experimental tasks, outcome measures, and methodological approaches. This variability limits comparability and prevents the identification of stable causal relationships between specific warning characteristics and behavioral outcomes. Accordingly, the patterns identified in this review should be interpreted as descriptive and exploratory rather than as evidence of the effectiveness of particular warning designs.

A third limitation concerns the coding of the outcome dimensions. These dimensions were not independently double-coded, and no inter-coder reliability coefficient was computed. However, the main coding decision was a low-inference binary judgment about whether each dimension had been assessed in a given study. This judgment was anchored to the explicit eligible and non-eligible indicators reported in Table 2. This procedure reduced, but did not eliminate, the scope for subjective interpretation.

A further limitation concerns the nature of the available evidence. Most included studies focus on short-term experimental outcomes measured during a single interaction or task session. They often rely on immediate behavioral indicators such as clicks, choices, response times, or classification accuracy. As a result, the review is constrained by the limited availability of longitudinal evidence on persistence, transfer, and stabilization of adaptive security behaviors over time. Trust calibration and risk regulation are also rarely assessed through standardized measures and are often inferred indirectly from behavioral responses. These limitations define the boundary of the present synthesis: the review maps how learning-relevant mechanisms are currently studied, but it does not estimate the effectiveness of specific warning designs.

Finally, no formal quality appraisal was conducted, in line with the aims of a scoping review, which maps the breadth of the literature rather than evaluating intervention effectiveness or excluding studies on methodological grounds. This means that limitations of individual studies were not systematically weighted in the synthesis.

4.3. Future Directions

Future studies should distinguish repeated exposure while the warning remains available from post-withdrawal maintenance, that is, persistence of safer behavior after the warning is removed. The descriptive patterns identified in this review define a set of testable priorities for future research. Future studies should examine whether warnings become more effective when they make risk interpretable, link alerts to concrete features of the user's current action, provide feedback on response accuracy, and adapt the level of guidance to user expertise.

More specifically, future research should test whether contextualized explanations help users understand not only that an action is risky, but why it is risky. It should also test whether feedback mechanisms improve trust calibration over time by informing users whether a warning reflected a genuine threat, whether their response was appropriate, or whether the alert was a false positive. Finally, future work should compare adaptive warnings with uniform warnings, because users with different expertise and prior exposure may need different levels of guidance.

Future experimental work should measure, alongside immediate compliance, comprehension of why an action is risky, alignment of reliance with alert reliability across true- and false-alert conditions, and retention of safer behavior across sessions and contexts. For example, studies could assess whether warnings help users and operators maintain effective decision-making under uncertainty, recover from inappropriate responses, or coordinate verification and reporting practices across repeated encounters with cyber risks.

These designs would allow researchers to distinguish immediate warning adherence from learning, calibrated reliance, and durable risk regulation.

5. Conclusions

This scoping review examined how cybersecurity warning systems have been empirically studied in relation to behavioral adaptation, trust calibration, and risk regulation, with particular attention to their potential role as safety-relevant learning mechanisms. As a focused scoping review of cybersecurity warning studies relevant to learning, this review examines a deliberately delimited corpus. It is therefore not intended to provide a comprehensive survey of the cybersecurity warning literature as a whole. The reviewed evidence indicates that warnings can influence immediate user behavior, but findings remain heterogeneous and largely limited to short-term responses within specific experimental settings. Current research is therefore better developed for assessing immediate warning response than for evaluating durable learning, calibrated reliance, or transferable risk-management strategies.

No single warning characteristic emerged as being consistently associated with safer behavior. Warning effects appear to depend on how users interpret, trust, and integrate risk signals within decision-making processes shaped by uncertainty, context, prior experience, and attentional demands. The review also identifies central gaps: comprehension is assessed infrequently, corrective feedback is largely absent, trust calibration is not formally measured, and longitudinal persistence or transfer is not evaluated.

Cybersecurity warnings should therefore be evaluated not only as mechanisms for interrupting risky actions, but also as components of socio-technical contingencies that may support learning, adaptive risk regulation, and more autonomous decision-making when feedback, repeated exposure, maintenance, and transfer are directly assessed. Future warning design should test whether contextualized explanations, corrective feedback, adaptive support, and explicit measures of trust calibration improve not only immediate compliance, but also the maintenance and transfer of safer behavior over time.

For safety practice, warnings should be embedded within organizational risk management rather than treated as isolated interface pop-ups. Warning policies should consider workload, interruption costs, false-positive costs, feedback loops, and the calibration of trust in automated detection. The learning-oriented framework proposed here provides a structured agenda for identifying which mechanisms are currently measured, which remain untested, and which design dimensions should guide the next generation of cybersecurity-warning studies.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/safety12040087/s1, Table S1: PRISMA 2020 Sc-R checklist.

Author Contributions: Conceptualization, E.C., L.A. and F.D.N.; methodology, E.C., L.A. and F.D.N.; formal analysis, E.C. and L.A.; investigation, E.C. and L.A.; data curation, E.C. and L.A.; writing—original draft preparation, E.C. and L.A.; writing—review and editing, E.C., L.A. and F.D.N.; visualization, E.C. and L.A.; supervision, F.D.N.; project administration, F.D.N. All authors contributed substantially to the work. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Veksler, V.D.; Buchler, N.; LaFleur, C.G.; Yu, M.S.; Lebiere, C.; Gonzalez, C. Cognitive models in cybersecurity: Learning from expert analysts and predicting attacker behavior. *Front. Psychol.* **2020**, *11*, 1049. [CrossRef] [PubMed]
2. Greco, F.; Desolda, G.; Buono, P.; Piccinno, A. Enhancing phishing defenses: The impact of timing and explanations in warnings for email clients. *Comput. Stand. Interfaces* **2025**, *93*, 103982. [CrossRef]
3. Chowdhury, N.H.; Adam, M.T.; Skinner, G. The impact of time pressure on cybersecurity behaviour: A systematic literature review. *Behav. Inf. Technol.* **2019**, *38*, 1290–1308. [CrossRef]
4. Di Nocera, F.; Tempestini, G.; Presaghi, F. Reliability and validity of the Cybersecurity Awareness INventory (CAIN). *Behav. Inf. Technol.* **2025**, *44*, 1417–1428. [CrossRef]
5. Chen, J.; Mishler, S.; Hu, B.; Li, N.; Proctor, R.W. The description-experience gap in the effect of warning reliability on user trust and performance in a phishing-detection context. *Int. J. Hum.-Comput. Stud.* **2018**, *119*, 35–47. [CrossRef]
6. Chen, J.; Mishler, S.; Hu, B. Automation error type and methods of communicating automation reliability affect trust and performance: An empirical study in the cyber domain. *IEEE Trans. Hum.-Mach. Syst.* **2021**, *51*, 463–473. [CrossRef]
7. Egelman, S.; Schechter, S. The importance of being earnest [in security warnings]. In *International Conference on Financial Cryptography and Data Security*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 52–59. [CrossRef]
8. Reeder, R.W.; Felt, A.P.; Consolvo, S.; Malkin, N.; Thompson, C.; Egelman, S. An experience sampling study of user reactions to browser warnings in the field. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–13. [CrossRef]
9. Sunshine, J.; Egelman, S.; Almuhimedi, H.; Atri, N.; Cranor, L.F. Crying wolf: An empirical study of ssl warning effectiveness. In Proceedings of the 18th USENIX Security Symposium, Montreal, QC, Canada, 12–14 August 2009; pp. 399–416.
10. Jones, K.S.; Lodinger, N.R.; Widlus, B.P.; Namin, A.S.; Hewett, R. Do warning message design recommendations address why non-experts do not protect themselves from cybersecurity threats? A review. *Int. J. Hum.-Comput. Interact.* **2021**, *37*, 1709–1719. [CrossRef]
11. Tempestini, G.; Rovira, E.; Pyke, A.; Di Nocera, F. The cybersecurity awareness inventory (CAIN): Early phases of development of a tool for assessing cybersecurity knowledge based on the ISO/IEC 27032. *J. Cybersecur. Priv.* **2023**, *3*, 61–75. [CrossRef]
12. Kelley, T.; Bertenthal, B.I. Attention and past behavior, not security knowledge, modulate users' decisions to login to insecure websites. *Inf. Comput. Secur.* **2016**, *24*, 164–176. [CrossRef]
13. Prebot, B.; Du, Y.; Gonzalez, C. Learning about simulated adversaries from human defenders using interactive cyber-defense games. *J. Cybersecur.* **2023**, *9*, tyad022. [CrossRef]
14. Gay, C.; Horowitz, B.; Elshaw, J.J.; Bobko, P.; Kim, I. Operator suspicion and human-machine team performance under mission scenarios of unmanned ground vehicle operation. *IEEE Access* **2019**, *7*, 36371–36379. [CrossRef]
15. Baltuttis, D.; Teubner, T. Effects of visual risk indicators on phishing detection behavior: An eye-tracking experiment. *Comput. Secur.* **2024**, *144*, 103940. [CrossRef]
16. Vance, A.; Eargle, D.; Kirwan, C.B.; Anderson, B.B.; Jenkins, J.L. The Fog of Warnings: How Non-Security-Related Notifications Diminish the Efficacy of Security Warnings. *MIS Q.* **2025**, *49*, 1357–1384. [CrossRef]
17. Arciulo, L.; Di Nocera, F. The Nudging Paradigm in Cybersecurity Research: A PRISMA-Based Systematic Review. *Information* **2026**, *17*, 264. [CrossRef]
18. Di Nocera, F. Usable security: A (re)definition and a research agenda. *Theor. Issues Ergon. Sci.* **2026**, *27*, 118–131. [CrossRef]
19. Amran, A.; Zaaba, Z.F.; Mahinderjit Singh, M.K. Habituation effects in computer security warning. *Inf. Secur. J. Glob. Perspect.* **2018**, *27*, 192–204. [CrossRef]
20. Di Nocera, F.; Tempestini, G.; Orsini, M. Usable security: A systematic literature review. *Information* **2023**, *14*, 641. [CrossRef]
21. Acquisti, A.; Adjerid, I.; Balebako, R.; Brandimarte, L.; Cranor, L.F.; Komanduri, S.; Leon, P.G.; Sadeh, N.; Schaub, F.; Sleeper, M.; et al. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Comput. Surv. (CSUR)* **2017**, *50*, 1–41. [CrossRef]
22. Rajarathnam, S.; Singh, V. Systematic literature review of cybersecurity and user experience. In Proceedings of the IEEE 2024 Cyber Awareness and Research Symposium (CARS), Grand Forks, ND, USA, 28–29 October 2024; pp. 1–9. [CrossRef]
23. Macmillan, N.A. Signal Detection Theory. In *Stevens' Handbook of Experimental Psychology: Methodology in Experimental Psychology*, 3rd ed.; Pashler, H., Wixted, J., Eds.; John Wiley & Sons: New York, NY, USA, 2002; Volume 4, pp. 43–90. [CrossRef]
24. Parasuraman, R.; Riley, V. Humans and Automation: Use, Misuse, Disuse, Abuse. *Hum. Factors* **1997**, *39*, 230–253. [CrossRef]
25. Parasuraman, R.; Hancock, P.A. Using Signal Detection Theory and Bayesian Analysis to Design Parameters for Automated Warning Systems. In *Automation Technology and Human Performance: Current Research and Trends*; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1999; pp. 63–67. [CrossRef]
26. Lee, J.D.; See, K.A. Trust in Automation: Designing for Appropriate Reliance. *Hum. Factors* **2004**, *46*, 50–80. [CrossRef] [PubMed]

27. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.J.; Horsley, T.; Weeks, L.; et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [CrossRef] [PubMed]
28. Maram, B.; Balija, S.; Sekhar, M.C.; Rao, L.G.; Raj, M.S.S.; Sathishkumar, V.E. XAI-Driven Security Systems: Enhancing Trust and Clarity in Automated Threat Detection and Response. In Proceedings of the IEEE 2nd International Conference on Intelligent Systems for Cybersecurity (ISCS), Gurugram, India, 14–15 November 2025; pp. 1–7. [CrossRef]
29. McRee, G.R. Improved detection and response via optimized alerts: Usability study. *J. Cybersecur. Priv.* **2022**, *2*, 379–401. [CrossRef]
30. Rasmussen, J.; Ehrlich, K.; Ross, S.; Kirk, S.; Gruen, D.; Patterson, J. Nimble cybersecurity incident management through visualization and defensible recommendations. In Proceedings of the Seventh International Symposium on Visualization for Cyber Security, Ottawa, ON, Canada, 14 September 2010; pp. 102–113. [CrossRef]
31. Bostan, A. Implicit learning with certificate warning messages on SSL web pages: What are they teaching? *Secur. Commun. Netw.* **2016**, *9*, 4295–4300. [CrossRef]
32. Blancaflor, E.; Deldacan, L.F.; Hunat, S.; Rivera, B.M.; Liberato, E.K. Ai-driven phishing detection: Combating cyber threats through homoglyph recognition and user awareness. In Proceedings of the 6th World Symposium on Software Engineering (WSSE), Kyoto, Japan, 13–15 September 2024; pp. 226–231. [CrossRef]
33. Datta, P.; Namin, A.S.; Jones, K.S.; Hewett, R. Warning users about cyber threats through sounds. *SN Appl. Sci.* **2021**, *3*, 714. [CrossRef]
34. Howell, C.; Maimon, D.; Muniz, C.; Kamar, E.; Berenblum, T. Engaging in cyber hygiene: The role of thoughtful decision-making and informational interventions. *Front. Psychol.* **2024**, *15*, 1372681. [CrossRef] [PubMed]
35. Stembert, N.; Padmos, A.; Bargh, M.S.; Choenni, S.; Jansen, F. A study of preventing email (spear) phishing by enabling human intelligence. In Proceedings of the IEEE European Intelligence and Security Informatics Conference, Manchester, UK, 7–9 September 2015; pp. 113–120. [CrossRef]
36. Zheng, S.Y.; Becker, I. Checking, nudging or scoring? Evaluating e-mail user security tools. In Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023), Anaheim, CA, USA, 6–8 August 2023; pp. 57–76. Available online: https://dl.acm.org/doi/10.5555/3632186.3632190 (accessed on 24 June 2026).
37. Roden, W.; Layman, L. Cry wolf: Toward an experimentation platform and dataset for human factors in cyber security analysis. In Proceedings of the ACM Southeast Conference, Online, 2–4 April 2020; pp. 264–267. [CrossRef]
38. Desolda, G.; Aneke, J.; Ardito, C.; Lanzilotti, R.; Costabile, M.F. Explanations in warning dialogs to help users defend against phishing attacks. *Int. J. Hum.-Comput. Stud.* **2023**, *176*, 103056. [CrossRef]
39. Sharevski, F.; Devine, A.; Pieroni, E.; Jachim, P. Phishing with malicious QR codes. In Proceedings of the European Symposium on Usable Security, Karlsruhe, Germany, 29–30 September 2022; pp. 160–171. [CrossRef]
40. Shava, F.B.; Van Greunen, D. Factors affecting user experience with security features: A case study of an academic institution in Namibia. In Proceedings of the IEEE Information Security for South Africa, Johannesburg, South Africa, 14–16 August 2013; pp. 1–8. [CrossRef]
41. Al-Hashem, N.; Saidi, A. The psychological aspect of cybersecurity: Understanding cyber threat perception and decision-making. *Int. J. Appl. Mach. Learn. Comput. Intell.* **2023**, *13*, 11–22. Available online: https://student.wp.odu.edu/ihenr002/wp-content/uploads/sites/35258/2024/04/ThePsychologicalAspectofCybersecurity.pdf (accessed on 24 June 2026).
42. Vance, A.; Jenkins, J.L.; Anderson, B.B.; Bjornn, D.K.; Kirwan, C.B. Tuning Out Security Warnings: A Longitudinal Examination of Habituation Through fMRI, Eye Tracking, and Field Experiments. *MIS Q.* **2018**, *42*, 355–380. [CrossRef]
43. Barbierato, E. Crying Wolf in Cyberspace: A Cybersecurity Dynamics Study of Alarm Fatigue Attacks. *Information* **2026**, *17*, 434. [CrossRef]
44. Sutton, J.; Wood, M.M. Opting Out: Over-Alerting and Warning Fatigue in the Era of Wireless Emergency Alerts. *J. Contingencies Crisis Manag.* **2025**, *33*, e70076. [CrossRef]
45. Almuhimedi, H.; Felt, A.P.; Reeder, R.W.; Consolvo, S. Your reputation precedes you: History, reputation, and the chrome malware warning. In Proceedings of the 10th Symposium on Usable Privacy and Security (SOUPS 2014), Menlo Park, CA, USA, 9–11 July 2014; pp. 113–128.
46. Roy, S.S.; Torres, C.; Nilizadeh, S. "Explain, Don't Just Warn!"—A Real-Time Framework for Generating Phishing Warnings with Contextual Cues. *arXiv* **2025**, arXiv:2505.06836. [CrossRef]