Information & Management 58 (2021) 103394

Contents lists available at ScienceDirect

Information & Management

journal homepage: www.elsevier.com/locate/im

Trust calibration of automated security IT artifacts: A multi-domain study of phishing-website detection tools

Yan Chen [a,*], Fatemeh Mariam Zahedi [b], Ahmed Abbasi [c], David Dobolyi [c]

[a] College of Business, Florida International University, 11200 S.W. 8th St., RB 203A, Miami, FL 33199, United States
[b] Sheldon B. Lubar School of Business, University of Wisconsin–Milwaukee, 3202 N Maryland Ave, Milwaukee, WI 53202, United States
[c] Mendoza College of Business, University of Notre Dame, 204 Mendoza College of Business, Notre Dame, IN 46556, United States

ARTICLE INFO

Keywords:
Trust calibration
Automated security IT
Phishing websites
Detection tools
Trust in IT

ABSTRACT

Phishing websites become a critical cybersecurity threat affecting individuals and organizations. Phishing-website detection tools are designed to protect users against such sites. Nevertheless, detection tools face serious user trust and suboptimal performance issues which require trust calibration to align trust with the tool's capabilities. We employ the theoretical framework of automation trust and reliance as a kernel theory to develop the trust calibration model for phishing-website detection tools. We test the model using a controlled lab experiment. The results of our analysis show that users' trust in detection tools can be calibrated by trust calibrators. Moreover, users' calibrated trust has significant consequences, including users' tool reliance, use, and performance against phishing websites.

1. Introduction

Phishing websites victimize millions of Internet users, exacting significant monetary losses and social costs for individuals and organizations [1–3]. An FBI announcement showed that phishing rendered $26 billion damage over a three-year period from 2016 to 2019 [4]. About $1.1 million per hour is lost to phishing attacks [5].

Phishing websites come in two forms: spoof and concocted. Spoof sites mimic existing, generally well-known websites to engage in identity theft or malware dissemination [6,7]. Concocted sites are fictional websites designed to conduct social engineering, fraudulent online advertising, or black-hat search engine optimization-based attacks for monetary gains or malware propagations. Both categories of phishing websites have serious implications for Internet users and organizations, such as damaging brand equity and increasing customer churn rates [6]. Concocted websites also frequently appear in top-ranked search results [8] and routinely disseminate malware to unsuspecting site visitors [9]. Phishing-website detection tools protect users against such sites.

These detection tools belong to a subcategory of IT called *automated security IT* and are defined as a type of security IT that uses certain mechanisms to automatically classify an event/objective as normal or malicious [10] while allowing users to make the final security decision [11]. There are many phishing-website detection tools, but reports indicate that users often ignore or disuse their advice [12,13]. A survey of Internet users found that 60 % of respondents do not use the web browsers' built-in phishing-website detection tools [14]. Many users rely solely on intuition to judge the credibility of a website despite the fact that spoof rates can be as high as 33 %–45 % when users rely on their own mental model [9,15,16]. While research shows that user accuracy in detecting phishing websites is much lower than the accuracy of the detection tools [1], the rate of ignoring certain types of warnings in some browsers (e.g., SSL warnings) can be as high as 60 % [17]. These results suggest that detection tools face serious trust issues in users. Addressing these issues demands a novel approach to investigate user trust vis-à-vis characteristics of detection tools.

Research shows that IT characteristics influence users' various perceptions, emotions, attitudes, and behaviors [18–21] (see Online Appendix A). Similarly, the characteristics of security detection tools have multiple influences and have been examined from multiple points of view, such as design of warning signals [22], neurophysiological impacts [23,24], and threat and coping appraisal perceptions [3].

One of the most important factors impacted by the IT characteristics is trust in IT. Information Systems (IS) research has extensively investigated trust in IT at both organizational and individual levels (e.g., [25–28]). However, the research has focused mainly on general-purpose IT and e-commerce—also referred to as positive IT [29]. Positive IT

[*] Corresponding author.
E-mail addresses: yachen@fiu.edu (Y. Chen), zahedi@uwm.edu (F.M. Zahedi), aabbasi@nd.edu (A. Abbasi), ddobolyi@nd.edu (D. Dobolyi).

https://doi.org/10.1016/j.im.2020.103394
Received 13 November 2019; Received in revised form 16 November 2020; Accepted 18 November 2020
Available online 25 November 2020
0378-7206/Published by Elsevier B.V.

strives for high levels of trust in order to increase user adoption and usage [20,30]. However, this is not necessarily the case for security detection tools. An "inappropriate" level of trust in detection tools can have negative financial and privacy consequences for users, leading to subsequent loss of their trust and use [13,31].

We address the issue of inappropriate trust by arguing that trust in such detection tools should be *calibrated* by closely aligning trust to the capability of the tool. Proper trust in detection tools requires matching trust with the tools' capability to identify phishing websites. Detection tools with high capabilities should enjoy higher levels of users' trust relative to those with low capabilities. Hence, we posit that calibrating users' trust in detection tools should play an important role in promoting protection against phishing websites.

Trust calibration is a concept developed in automation research. It is defined as the process of creating correspondence between the extent of users' trust in an automated tool and the capability of that tool [32–34]. C*alibrated trust* is the level of trust that matches the capability of the tool—the level of user trust after gaining knowledge about the tool's capability [33]. In spite of its importance, IS research has not addressed inappropriate trust and trust calibration in automated security IT. In practice, while automation research has reported that providing accurate and trustworthy information on automation capability and process (e.g., via display or interface design) influences trust calibration [35], detection tools do not provide such information. Our paper addresses this research gap with the focus on phishing-website detection tools as a type of automated security IT that allows users to make the final decision.

This paper takes a theoretical approach to identify the salient factors (or trust calibrators) in calibrating trust in phishing-website detection tools and the outcomes of calibrated trust in such tools for individual users. Therefore, the research questions in this paper are as follows—-What are the antecedents of calibrated trust in phishing-website detection tools for individual users? What are the consequences of calibrated trust in these tools for individual users?

To address these research questions, we employ the theoretical framework of automation trust and reliance (ATR) [33] as a kernel theory to develop the trust calibration model for phishing-website detection tools (referred to as the TC model for brevity). We test the TC model using a controlled lab experiment. The results uncover the importance of detection rate and outcome severity in calibrating trust in phishing-website detection tools and the central role of calibrated trust in users' higher reliance, use, and protection against phishing websites. We discuss the theoretical and practical implications of our work for individuals, organizations, and security IT tool designers.

2. Literature review

Automated security IT artifacts carry out partially or fully automatic protective functions such as detecting, deterring, disabling, or eliminating the security threats a user could encounter when using IT. As such, phishing-website detection tools fall under automation, defined as a technology that "accomplishes (partially or fully) a function that was previously, or conceivably could be, carried out (partially or fully) by a human operator" ([36], p. 287). Different from automated security IT artifacts with no user choice (e.g., firewalls), phishing-website detection tools give users the choice of accepting or rejecting their advices [11]. The automatic nature of security detection tools with user choice plays a critical role in their trust calibration, and places the study of such calibration within the domain of automated security IT.

While trust in general-purpose IT is viewed as desirable—the higher the trust, the more usage [20,30], this is not necessarily the case for automated security IT that end users can ignore its advice. For such IT, only an "appropriate" level of trust is desirable in order to protect users from harms and losses. Overtrust may cause abuse—people overly rely on it and incur losses due to its errors in detecting threats. Undertrust leads to disuse—people reject its capabilities and incur losses due to ignoring its advice. This is shown to be true for all automated tools that are imperfect [33,34,37–39]. Therefore, trust calibration is critical for a proper level of trust. Without calibrated trust, phishing-website detection tools demonstrate disappointing performance or disuse [13,40].

Past research on trust in IT has focused on general-purpose IT and sought to increase trust based on the assumption that the higher the trust, the higher the usage (see the summary of our literature review in Table B-1 of Online Appendix B). Another assumption is that IT has social presence and users have opportunities to interact with IT to form or increase their trust [20]. As such, trustworthiness beliefs are defined as trust antecedents, and these beliefs are critical in forming and increasing trust. Research has borrowed ability, integrity, and benevolence as trustworthiness beliefs from interpersonal trust studies to investigate trust in IT, including more recent studies on trust in artificial intelligence IT (e.g. [41],), and found that such beliefs are shown to be significant in increasing trust [27,41]. When the IT artifact has social presence (such as recommendation agents that act as human agents and Facebook with its features for interpersonal interactions), the literature suggests that interpersonal trustworthiness beliefs (ability, integrity, and benevolence) explain users' perceptions and behaviors [20,27,41].

However, the role of such trustworthiness beliefs has come under question when IT artifacts lack social presence (e.g., MS Access) [20]. Research suggests that for IT artifacts without social presence, trust should be based on system features [20,26]. Studies of trust in artifacts without social presence is scarce. Research in trust in automated security IT is almost non-existent. One exception is a study on antivirus software [42], which reports performance, predictability, and subjective norms as predictors of trust in and satisfaction with the artifact and argues that some of these predictors partially overlap interpersonal trustworthiness beliefs.

Automated security IT that advises users is different by nature from general-purpose IT (see the summary in Table 1). In addition to its lack of social presence, it runs automatically in the background with minimal human-computer interactions and is not the primary focus of users' activities. As a result, users tend to ignore security tools' warnings for the expediency of accomplishing their primary task—access to their intended websites. Even when they follow the advice of the tool, they do so with little knowledge about and feedback from the tool. In other words, users have no detailed information about or interactions with the

Table 1
Characteristics of Automated Security IT Compared to General IT.

Automated Security IT with User Choice[a]	General IT
Automated security IT lacks social presence.	General IT has a certain level of social presence.
Automated security IT deals with security tasks that are secondary to end-users [15,44]. When detecting fraud, security IT interferes with users' primary tasks. Consequently, users are less willing to focus on or attend to the advice of security IT.	General IT is a part of users' primary tasks, on which users focus and pay full attention.
Automated security IT works in the background, hidden from users. As such, characteristics of security IT are invisible to users. Consequently, users make security decisions with little knowledge about the tool [44].	General IT interacts with users via human-computer interactions. IT characteristics are more visible to users.
Automated security IT provides security, which is an abstract concept to users [43,44].	General IT assists users to improve performance, efficiency, and productivity of the task, which are concrete gains to users.
Automated security IT has persistent adversaries who try to undermine security tools' performance and victimize their users.	General IT does not have adversaries, who actively and persistently try to undermine the tools' performance.

[a] In this study, we study phishing-website detection tools as exemplars of automated security IT that allows users the choice of accepting or rejecting its advice.

tool to calibrate their trust, so they end up making security-related decisions in the dark. Moreover, the "security" such tools provide is an abstract concept for users [43,44] as security is invisible and its desirable consequence is a non-event. "The reward for being more secure is that nothing bad happens" ([43], p. 37). More importantly, such tools deal with changing and morphing adversaries who try to undermine tools' performance and victimize users. In summary, these unique characteristics demand a fresh perspective and theory-based analysis on trust in automated security IT with user choice in general and phishing-website detection tools in particular.

Trust calibration in automation has been studied (as reported in Table B-2 of Online Appendix B). As shown in those studies, various characteristics and performance metrics can be used to calibrate trust. Trust calibration is one of the most important design strategies that leads to proper reliance and human-automation performance [32–34,38,39,45]. Automation literature reports that trust in automation is turbulent, fragile, and tentative. It must be properly and promptly calibrated by salient calibrators that inform users about the key characteristics of the automated tools. Automated security artifacts too have various characteristics and form a category of automation. As such, trust in these artifacts needs to be calibrated.

Research in the fields such as automation and human factors has identified some salient calibrators—"error rate" as a calibrator for an automated route planner [46], "accuracy" for an automated screener [47], and "reliability" for systems such as vehicle control system, automated signaling system, automatic decision aids system for detecting infight icing events [38,39,45,48,49]. Research in these fields has taken a focus on mechanistic approaches (e.g., experiments) and paid less attention to theory building and development. Additionally, IS research has overlooked the concept of trust calibration and its significance along with trust calibrators and their effects especially in the context of automated security IT with user choice. One exception is a study by Chen et al. [40]. The study examined differences in trust by manipulating the calibrators (e.g., reliability and feedback) of a phishing-email detection, without offering much theorization for its work.

In this study, we focus on phishing-website detections tools (referred to as detectors) as exemplars of automated security IT that allows users the choice for accepting or rejecting its advice. We rely on a theoretical framework—the ATR framework—to identify salient trust calibrators and to examine the consequences of calibrated trust in detectors. Hence, the critical components of ATR and their relationships guide the conceptualization of our research model. As trust calibration requires knowledge of trust calibrators and repeated use of the detector, we test our conceptualized model with data obtained from a complex controlled experimental design with repeated observations and exposures to trust calibrators.

3. Theoretical framework

IS research on trust in IT has relied on a variety of theories such as trust beliefs (ability, integrity, and benevolence) (e.g., [28]), expectation disconfirmation theory (e.g. [20,42],), and theory of reasoned action (e.g. [27],). Phishing-website detections tools are highly automated, work behind the scenes, and need trust calibration—features that trust theories in prior research do not address. We use automation trust and reliance (ATR) as our theoretical framework because ATR focuses on the automated nature of security IT, draws a clear distinction between trust in automation and trust in humans, and provides an integrated perspective on trust calibration. We rely on the key components of ATR to identify the salient characteristics of phishing-website detection tools (as a specific category of automated security IT). According to ATR, users need to be informed about these characteristics for trust calibration process [33].

ATR has three main components, as shown in Fig. 1. The first component consists of the salient characteristics of the automation that calibrate trust. Calibrated trust is the result of trust evolution with the repeated use of the automation. This calibration depends on informing users of the salient characteristics (ex. via displays or prompts). The second main component of ATR consists of the outcomes of calibrated trust, such as intention formation and proper reliance on the automation [33]. The third component of ATR is contexts, including individual, organizational, and environmental contexts [33].

In the first component, ATR identifies three characteristics of automation: *performance*, *process*, and *purpose*—called trust calibrators in this study. ATR argues that trust calibrators in automation are the counterparts of trust beliefs (ability, integrity, and benevolence) and that they become the antecedents of trust [33]. *Performance* describes what the automated tool can do. It demonstrates the ability of the automation including predictability, accuracy, and reliability. *Process* reveals the underlining mechanisms and operations of the automation. Process as an antecedent of trust shows the openness and integrity of the

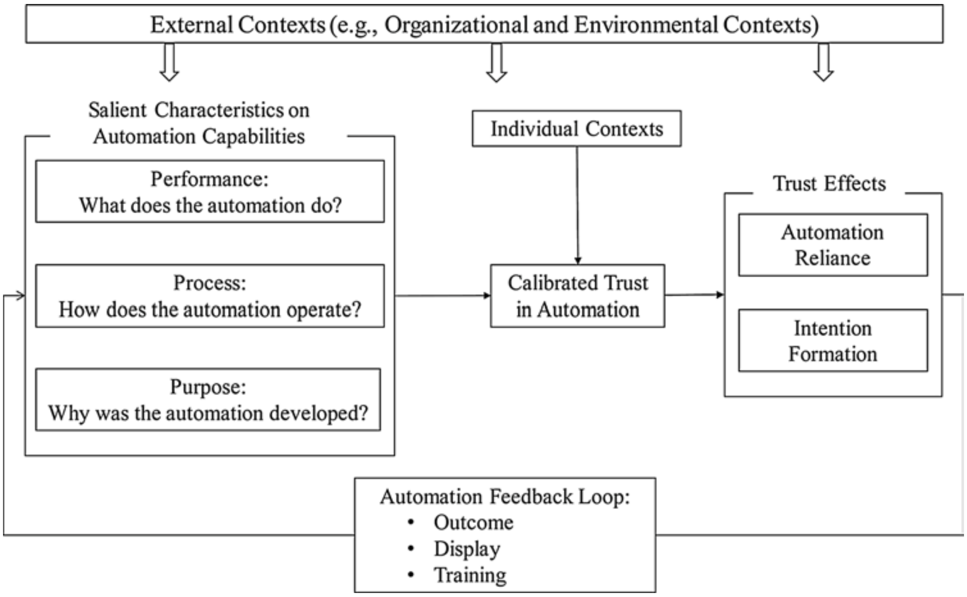

External Contexts (e.g., Organizational and Environmental Contexts)
Salient Characteristics on Automation Capabilities
Performance: What does the automation do?
Process: How does the automation operate?
Purpose: Why was the automation developed?
Individual Contexts
Calibrated Trust in Automation
Trust Effects
Automation Reliance
Intention Formation
Automation Feedback Loop:
• Outcome
• Display
• Training

Fig. 1. The Framework of Automation Trust and Reliance (ATR) (Adapted from Lee and See [33]).

automation [33]. *Purpose* explains why the automation exists and reflects the developer's intent and motivation. Purpose as a basis for trust shows how the automation intends to address users' needs [33]. ATR empathizes the importance of displaying the key characteristics of the automation to calibrate trust.

The second component we draw from the ATR framework is the outcomes of calibrated trust, which include reliance, performance, and intention aligned with automation capability. Reliance refers to how users behaviorally depend on automatic aids. Performance shows the results of using the automation. Intention projects use in the near future.

The third component we draw from the ATR framework is the influence of contexts in trust calibration, such as individual, organizational, and environmental contexts [33]. Individual differences such as habit and past experience may influence trust in automation. Organizational contexts such as reputation, gossip, organizational structures, values, and norms play a role in trust formation. Environmental variables, such as the environment in which the automation is used, also impact trust calibration [33].

We rely on the three ATR components to formulate our model—(i) characteristics of automation as antecedents of calibrated trust; (ii) outcomes of calibrated trust including reliance and use; and (iii) automation contexts—individual and environmental. Organizational context was not used as the unit of analysis in this study is at the individual level.

ATR posits the dynamic of trust calibration in automation—trust evolves through this process in which the experience of using the automation provides feedback for trust calibration [33]. Following research in the fields of automation and human factors [40,46,50,51], we incorporate the dynamic of trust calibration in the process trust calibration through multiple experiences of using the security tool in two distinct domains (health and finance). We measure participants' calibrated trust as their responses to the trust calibrators of the phishing-website detection tool after multiple experiences of using the tool in the experiment.

4. Model of trust calibration for phishing-website detection tools

The critical components and their relationships in ATR form the theoretical foundation on which we conceptualize the trust calibration model for automated security IT in the context of phishing-website detection tools (the TC model) (see Fig. 2). Briefly, the TC model shows the influence of trust calibrators on calibrated trust (H1-H3). H4-H7 capture the ATR-based outcomes of calibrated trust. The environmental context is captured through the moderation of domain type. Individual contextual factors are represented as control variables. Currently, the TC model focuses on how individual users respond to trust stimuli (i.e., calibrators) of the detector.

4.1. Performance: detector accuracy→calibrated trust (H1)

According to ATR, *performance* relates to what an automated tool does, including how reliably it can fulfill its expected purpose [33]. An important aspect of reliability is accuracy. Prior research indeed showed the effectiveness of accuracy as a performance-based trust calibrator [37,52]. Accuracy-related assessment on automation was also found to play a critical role in shaping or reshaping users' trust in automation [34]. Users are less likely to ignore a highly-accurate automation aid when they are informed of the details regarding the aid's performance [52].

In reality, automated tools often cannot consistently produce perfect results, and users are keenly aware of this discrepancy. In certain contexts, users will discount automated tools despite high accuracy because of their misjudgment of the discrepancy [34,53]. In the context of automated security IT, such misjudgment could happen more often as security is a secondary, background task to users [15] and security tools generally provide very little performance information to users [43]. With little or no information to draw upon, users tend to make a quick, even impatient, judgment about performance. Explicit detection accuracy information allows users to make a more accurate judgment about the discrepancy and thus calibrate their trust.

In terms of phishing, the detection rates of commonly used state-of-the-art detection tools currently range from approximately 60%–95% [1]. For detectors in some popular browsers, given the advances in detection accuracy, we would expect very low click-through rates [17]. However, this is not the case [17]. We argue that users need to know the tool's accuracy. When users are exposed to information about the detector's accuracy, they can calibrate their trust accordingly. Hence,

H1. Detector accuracy is positively associated with users' calibrated trust in the detector.

4.2. Process: detector run-time→calibrated trust (H2)

The second trust calibrator in the ATR framework is *process*, which relates to how an automated tool functions [33]. A critical aspect of process is transparency of the tool's mechanisms to users [33]. Although

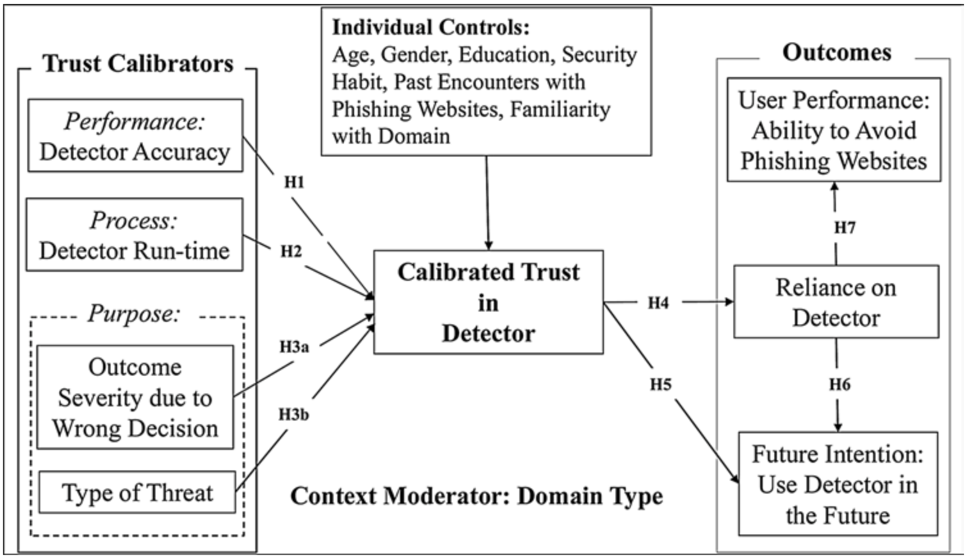

Trust Calibrators
Performance:
Detector Accuracy
Process:
Detector Run-time
Purpose:
Outcome Severity due to Wrong Decision
Type of Threat
Individual Controls:
Age, Gender, Education, Security Habit, Past Encounters with Phishing Websites, Familiarity with Domain
H1
H2
H3a
H3b
Calibrated Trust in Detector
H4
H5
Outcomes
User Performance: Ability to Avoid Phishing Websites
H7
Reliance on Detector
H6
Future Intention: Use Detector in the Future
Context Moderator: Domain Type

Fig. 2. Trust Calibration Model for Phishing-Website Detection Tools.

tools typically do not provide visible information regarding the details of the underlying analytical methods, users are often keenly aware of the run-time, particularly when the tool is perceived as slow. Run-time metrics are commonly used to assess credibility and performance for a wide range of automated tools, including phishing-websites detectors [7,54]. Nevertheless, run-time is often shown on the interface as a progress bar or circle that provides a symbolic indicator of process and visual feedback that the system is running and operating on the task. Improving run-time is a well-established way to improve user experience in terms of trust and satisfaction [55].

Phishing-website detection tools use sophisticated machine learning algorithms, pattern matching, and various large datasets in their detection process. However, users do not have the expertise and time to assess details of tools' methods. Detection tools often use a progress bar to indicate run-time, and thus run-time is the only aspect of a tool's detection process that users directly experience and feel its impact when accessing websites. When the process is slow, users become annoyed and frustrated because the detector is a secondary operation working behind the scenes. Its slow operation can hinder users' primary tasks [56]. Slow run-time could even cause users to suspect that the system's security is compromised and thus not trustworthy [55,57]. The prolonged interruption (a long run-time) challenges users' patience and potentially implies process inferiority to users of the tool. Hence,

H2. Shorter detector run-time is positively associated with users' calibrated trust in the detector.

4.3. Purpose: outcome severity due to wrong decision→calibrated trust (H3a)

The third trust calibrator in the ATR framework is *purpose*, which relates to why an automated tool exists [33]. A main purpose of security IT is to help users prevent and eliminate damaging consequences from security threats [29]. If a security tool fails to accurately alert a user about a potential security threat, it fails to fulfill its intended purpose. Automation research shows that the potential outcomes from such failure are often costly and sometime even disastrous, leading to a decline in trust [34,58]. Automation research also suggests that users tend to attribute the damage caused by their wrong decision to the automation, leading to trust reduction, even though they are ultimately responsible for the decision whether to heed the automation's warning [59].

In our research context, the purpose of a phishing-website detector is to protect users from unwittingly visiting phishing sites. Failure to do so would result in users' exposure to significant personal, professional, and/or financial risks such as identity theft or financial loss. If a phishing-website detector fails to accurately alert a user about a potential phishing site, it fails to fulfill its intended purpose. When the user follows the detector's incorrect advice, the outcomes can be costly and thus not easily forgotten or forgiven [1]. In fact, a single wrong decision can cause a long-lasting, biased view about the tool [60].

Research has found that the cost of decision error influences users' perceptions of a detection tool [3]. When the severity of outcomes due to errors of an automated system increases, users' trust in the system declines [58,61]. Applied to phishing-website detectors, we argue that if the outcome of a wrong decision is more severe, users form lower trust in the detector. Hence,

H3a. Outcome severity due to wrong decision is negatively associated with users' calibrated trust in the detector.

4.4. Purpose: type of threat→calibrated trust (H3b)

ATR argues that gaining a clearer understanding of the specific goal an automated tool is designed to achieve may lead users to place more proper trust in the tool [33]. Purpose-based information informs users about "the specific problem that the automation might have been designed to solve, as well as lower-level objectives that the automation was designed to meet" ([62], p. 3). Thus, a deeper understanding of an automated tool's purpose allows users to better determine if the tool meets their goals. This also allows users to foresee situations in which the tool might understandably fail [33,62]. As a result, users are able to place more appropriate trust in the tool [33,62].

In terms of phishing-website detectors, although all automated detectors are to detect and protect users from visiting phishing websites, some detectors have a more limited purpose. For example, the purpose of eBay's Account Guard toolbar is to detect spoof websites mimicking eBay. Other tools such as AZProtect and browsers' built-in anti-phishing tools are multi-purpose and designed to detect both spoof and concocted websites in all domains [1,13]. Another example is that if a tool is designed to detect phishing websites in a specific language, it is unable to detect such phishing websites in other languages because it has no capabilities to properly parse other languages [6]. In general, many existing detectors focus on detecting a single type of phishing site, and their detection capabilities across both concocted and spoof sites are different [7].

Thus, the type of threat a detector handles (i.e., concocted and/or spoof) provides information about the detector's specific purpose and the scope of its detection capabilities. As concocted and spoof sites use different deceptive tactics to defraud, users behave differently toward each type of threat and the detector targeting it [7,63]. When facing a concocted site, users are often influenced by its aesthetic and professional appearance; in dealing with a spoof site, they judge its legitimacy based on their past experience with the legitimate counterpart. Providing the information of threat types affords users a deeper understanding of the detector and allows them to rationally adjust their expectations regarding the detector [7], as well as adjust their detection strategies based on the type of the phishing website they encounter. We argue that such understanding enables users to better calibrate their trust in the tool. Hence,

H3b. Type of threat is associated with users' calibrated trust in the detector.

4.5. Outcomes: calibrated trust→reliance on the detector (H4)

ATR argues that once the automated advisor is adopted, reliance on its advice depends on the extent of user trust [33]. Calibrated trust "guides reliance when complexity and unanticipated situations make a complete understanding of the automation impractical" ([33], p. 50). In other words, reliance is a behavior outcome of trust under uncertainty, making trust an antecedent of reliance on automation [62,64].

It has been observed that people often turn to manual control to reduce their reliance on the automatic system if they do not trust it [33, 65,66]. In contrast, excessive, misplaced trust in automation may lead to over-reliance, which in turn leads to complacency, decreased vigilance, and less effort in monitoring automation performance. These findings show that calibrated trust is needed for an appropriate level of reliance [67–69].

In the context of phishing website detection, a lack of trust in a detection tool can significantly reduce users' reliance on its advice, causing them to turn to their own judgment when assessing the legitimacy of a website. This results in inadequate performance in detection of phishing websites [13]. Following this logic, we hypothesize,

H4. Users' calibrated trust in the detector is positively associated with their reliance on the detector's advice.

4.6. Outcomes: calibrated trust→future use (H5)

According to the ATR framework, another outcome of trust in automation is intention to use [33]. In the IS field, abundant empirical evidence has supported the relationship between trust and intentions in various contexts (e.g. [19,21,27,30,70],). For example, trust has been found to be a predictor of technology acceptance intentions [19], and trust in the IT artifact has also been found to be associated with intention to use the artifact [27,70]. Given that research has extensively documented the relationship between trust and intention to use, we argue that this relationship can also be applied to the current research context. Hence,

H5. Users' calibrated trust in the detector is positively associated with their intention to use the detector in the future.

4.7. Outcomes: reliance on the detector→future use (H6)

ATR posits that reliance is the level of users' behavioral dependence on the automation aid [33]. Such dependence exhibits inertia [71]. This means that present reliance and positive experiences with the automation can result in an intention to continue using it in the near future. Prior IS research has also demonstrated the link between current behaviors and continuance intentions [72]. Additionally, reliance is based on trust beliefs surrounding automation performance, openness and integrity in process, and automation intent and motivation [33]. As long as these beliefs remain unbroken, reliance on automation will remain strong and use will continue [33,34].

In the current research context, reliance is built upon calibrated trust that matches the capabilities of the detector. Thus, we argue that reliance will demonstrate inertia in that present positive interactions with the detector are a motivation of intention to use in the future. In other words, users tend to maintain their intention to use the detector if current reliance on the detector is properly established. Following this logic, we hypothesize,

H6. Users' reliance on the detector's advice is positively associated with their intention to use the detector in the future.

4.8. Outcomes: reliance on the detector→user performance (H7)

ATR argues that proper reliance determined by calibrated trust is the key to improving user performance. Prior research has shown that humans are remarkably poor in detecting deceptions. In some circumstances, professionals and novices alike can achieve only slightly better detection accuracy than that of flipping a coin [65,73]. As such, when automated tools have relatively high accuracy, humans are often the weak link in making correct decisions [11,68]. Their reliance on such tools could lead to better user performance in avoiding deception [11].

Phishing websites use deceptive tactics to lure in users by manipulating and misrepresenting cues that are present in many legitimate websites [13,63]. For instance, the visual appearance of a website, which can be easily manipulated, has been reported to be a prominent factor impacting users' credibility judgment about the website [13]. Not surprisingly, users who rely solely on their own judgment and abilities routinely fail to identify 35%–45% of phishing websites encountered, with misclassification rates as high as 70 % in some instances [15,74]. In contrast, state-of-the-art detection tools' misclassification rates for phishing websites are only roughly 10 % [1,7]. Thus, relying on the tool's advice results in better user performance [17]. Therefore, under the current research context, we expect that high reliance (resulting from a calibrated trust) leads to users' higher ability to avoid phishing websites. Hence,

H7. Users' reliance on the detector is positively associated with users' performance in terms of their ability to avoid phishing websites.

4.9. Contextual controls

ATR emphasizes the importance of the contextual controls in which trust is calibrated. Following this framework, we examine the environmental context for the phishing-website detectors by moderating domain type. Research has reported that context and domain play an important role in the implementation of behavioral theories [75]. For example, it is shown that domain has a significant effect on people's disclosure of their private information online [76]. For this reason, we chose to test the model in two domains—online pharmacy and online banking. Phishing attacks are prevalent in both domains, causing users to suffer identity theft and monetary losses [77]. The two domains have distinct security risks. Purchasing drugs from online pharmacies carries a high risk of encountering concocted online pharmacies and counterfeit products [77,78]. Online banking websites face more spoof attacks that attempt to directly defraud victims for financial gain [6].

Finally, ATR recognizes the importance of controlling for individual contextual factors that influence calibrated trust [33]. Thus, we include age, gender, education, security habit, and past encounters with phishing websites as the individual context.

5. Research methodology

We conducted a between-subject controlled lab experiment using a phishing-website detection tool (detector). The experiment consisted of a 2 (threat domain: bank vs. pharmacy) x 2 (type of threat: spoof vs. concocted) x 2 (accuracy of detector: high [90 %] vs. low [60 %]) x 2 (run-time of detector: fast [1 s] vs. slow [4 s]) x 2 (outcome severity due to wrong decision: high [$10] vs. low [$1]) full-factorial design with a total of 32 conditions.

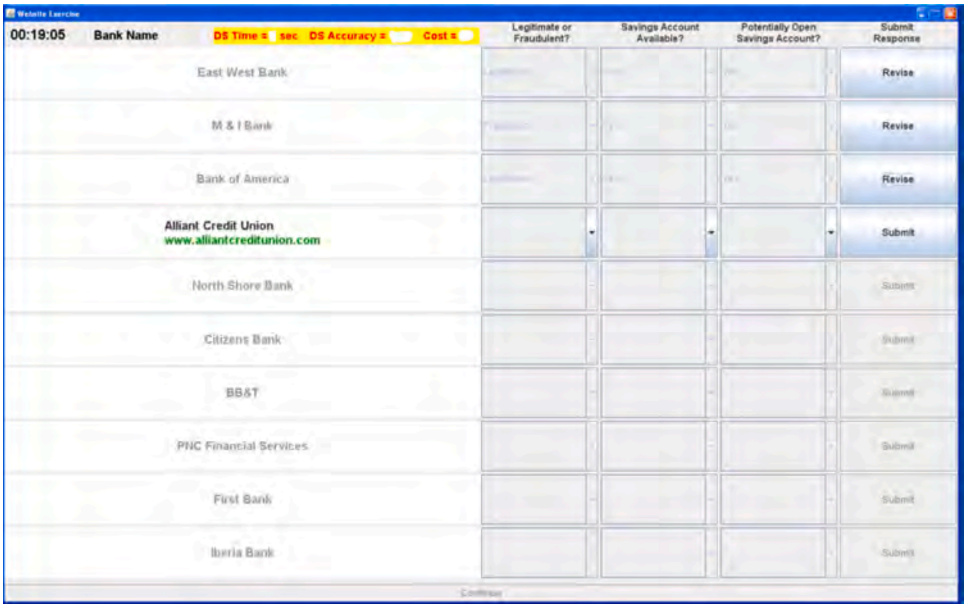
We created an inventory that contains the clones of 15 spoof, 15 concocted, and 15 legitimate websites for each domain. We collected phishing websites from reputable sources (e.g., LegitScript, PhishTank) and legitimate websites using a spidering program that preserved the original link structure, content, and images of the websites of the legitimate companies. To avoid company-size bias, the inventory included an equal number of large, medium, and small companies for the 15 legitimate websites.

Before the experiment, participants were informed of the experimental procedures and trained about the key concepts such as threat types (spoof vs. concocted) and detection accuracy (the percentage of all websites that are correctly classified) so that they would have a clear understanding of the goal and purpose of the detector and the type of threat they could encounter. Participants then completed a pre-experiment survey about their past experiences, security habits, and other relevant questions. To simulate a real condition, participants received money or course credit based on their security protection performance of avoiding phishing websites.

In the experiment, participants were randomly assigned five legitimate and five phishing (either spoof or concocted) websites and asked to perform a task according to their assigned domain (either the online pharmacy domain or the online bank domain). Hence, they had ten opportunities (10 trials) to use the detector and calibrate their trust in the detector.

ATR posits that exposure to salient trust calibrators is critical in trust calibration [33]. Research has shown that explicitly providing real-time confidence levels helped calibrate users' trust in automation aids [37]. Thus, in the experiment, participants were explicitly informed of the accuracy, run-time, and outcome severity of a wrong decision by a display on the top of screen during the entire period of the experiment. Fig. 3 shows the interface of the experiment with the links to the 10 assigned websites along with the information of the trust calibrators on the top bar (see the yellow highlights in Fig. 3).

When participants click a link, the detector shows the progress bar for 1 or 4 s and then delivers the detection outcome. For the run-time, we used a progress bar to visually show the time it took to run the detector in the background. While the lab-experiment method permits designs that deviate from real experiences, we chose to preserve the realism in our experimental design by using a progress bar. The progress bar communicated the run-time information in a familiar and easily understandable way to the participants without distracting them from their main tasks. Fig. 4-Panel A shows an example of the warning to

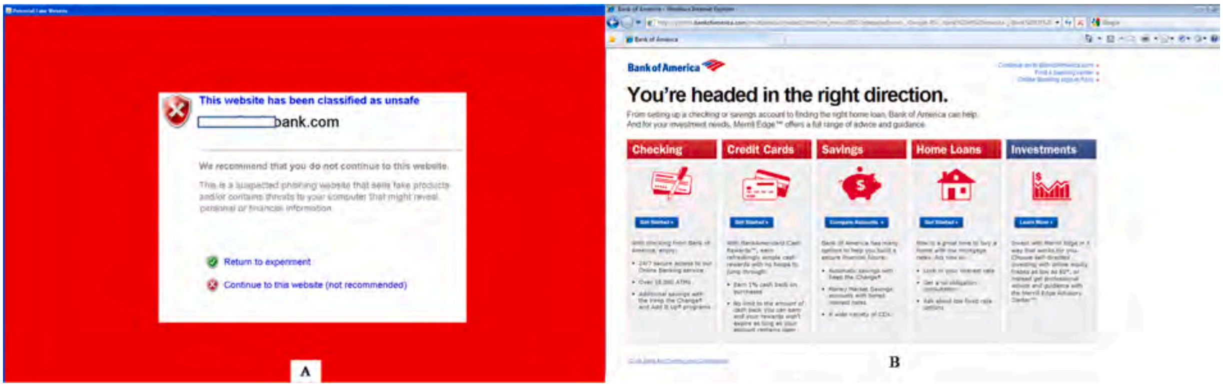


Website Exercise

00:19:05 Bank Name	DS Time = sec DS Accuracy = Cost =	Legitimate or Fraudulent?	Savings Account Available?	Potentially Open Savings Account?	Submit Response
East West Bank					Revise
M & I Bank					Revise
Bank of America					Revise
Alliant Credit Union www.alliantcreditunion.com					Submit
North Shore Bank					Submit
Citizens Bank					Submit
BB&T					Submit
PNC Financial Services					Submit
First Bank					Submit
Iberia Bank					Submit

Continue

Fig. 3. Experiment Interface with Information on Trust Calibrators.



This website has been classified as unsafe

[redacted]bank.com

We recommend that you do not continue to this website.

This is a suspected phishing website that sells fake products and/or contains threats to your computer that might reveal personal or financial information.

Return to experiment

Continue to this website (not recommended)

Bank of America

You're headed in the right direction.

From setting up a checking or savings account to finding the right home loan, Bank of America can help. And for your investment needs, Merrill Edge™ offers a full range of advice and guidance.

Fig. 4. Detection Outcome Examples.

participants when a website was detected as a phishing site, while Fig. 4-Panel B shows that participants directly accessed the webpage without any warning block when a website is classified as legitimate. In doing so, we ensured participants' trust was calibrated by the trust calibrator information displayed on Fig. 3 and by their repeated interactions with the tool.

We informed participants of their detection performance after the participants finished with the 10 repeated trials and before answering the post-experiment questions. We used this design because, in reality, security detection tools are not 100 % accurate and cannot give immediate feedback on the correctness of users' decisions. People also do not see the consequences (e.g., identity theft) of detection errors right away. Additionally, if we provided performance feedback after each trial to participants, they could have guessed the correctness of the detection outcome during the remaining trials.

All participants performed a domain-related task. In the online pharmacy domain, the experimental task was to purchase Rogaine, a popular over-the-counter hair restoration drug. This product was chosen because it is familiar and carried by most online pharmacies. Moreover, counterfeit Rogaine is often sold by phishing websites. In the online bank domain, the experimental task was to open an online saving account, which is a basic function provided by most online banks. This task is relevant as providing personal and financial information to a phishing bank website poses great risk of financial loss and identity theft. Regardless of domain, participants had to make a series of decisions about each website they encountered, including whether to visit or browse the website, whether they considered the website legitimate or phishing and whether they would transact with the website (see Fig. 3). Visiting and browsing behavior during the experiment was measured using web analytics software that tracked users' clicks. This experiment design also allowed participants to have multiple interactions with the detector for appropriate trust calibration.

All participants started with a cash box of $100 (hypothetical money). Every time they made a wrong decision, they would lose money ($1 or $10, depending on threat severity). Visiting a phishing website was also penalized as a wrong decision as visiting such a website carries the risk of being victimized by malware and other security threats [60]. Participants were compensated with a uniform base plus extra compensation depending on the money left in their cash box. This performance-dependent compensation was designed to increase the motivation of participants to perform well in the experiment. A final performance score for each participant was computed based on all their decisions regarding their 10 assigned websites. Based on the final scores, participants were paid a minimum of $10 and a maximum of $30 or extra credit for their participation (depending upon their preferences). The experiment was conducted using a Java-based software tool specifically developed for this study. Online Appendix C provides additional details on the experiment protocol and the role of the Java tool in administrating the protocol.

After the experiment, participants were informed of their detection performance. They then completed a post-experiment survey consisting of manipulation checks and perceptual questions.

Overall, this full-factorial, controlled lab experiment design allowed us to directly establish the causality between trust calibrators and calibrated trust. We randomly varied the levels of tool capability (and other calibrators) assigned to each participant, while fully controlling for other sources of variation. Each participant learned about the trust calibrators of the tool assigned to them from the start, made his/her detection dections based on what he/she was informed, and was compensated depending on his/her security protection performance. We measured calibrated trust as well as the outcome variables at the end to capture the causality between trust calibrators and calibrated trust and its consequences.

6. Scale development and data collection

To ensure validity whenever possible, measurement scales of the constructs in the TC model were adopted from existing literature. In addition, all items were converted to semantic differential scales to ensure content validity and reduce the threat of common method bias [79,80]. The items for calibrated trust in the detector were self-developed based on Bansal et al. [81]Gefen et al. [19], and Madsen and Gregor [82]. In the fields such as automation and human factors, most studies on trust calibration use a 5- or 7-point scale [40,46,51] or short version of 3-item construct adopted from the trust literature to measure calibrated trust (e.g. [38],). This is to reduce frustration of the participants in the face of multiple trials in the studies. Those studies also suggest that such a measure is able to capture calibrated trust after subjects interact with the stimuli/calibrators of the automation manipulated in research [50]. In the IS field, three dimensions of competence/functionality, integrity/reliability, and benevolence/helpfulness are used to measure trust beliefs in various contexts. Considering the practice in these research fields, we used three items to measure calibrated trust. In detail, the item of “not reliable at all/very reliable for sure” is based on the reliability dimension for trust in [82], the item of “not dependable at all/very dependable for sure” is based on the technical competence dimension for trust in Madsen and Gregor [82] and the opportunistic/dependable dimension for trust in Gefen et al. [19] and Bansal et al. [81], and the item of “not trustworthy at all/very trustworthy for sure” is based on honesty/benevolence dimension in Gefen et al. [19] and Bansal et al. [81]. Intention items to use the detector in the future were adapted from Davis et al. [83]. Finally, the items for reliance on the detector were adapted from Davis et al. [83] and Venkatesh et al. [84]. All the latent constructs are reflective.

Participants’ ability to detect phishing websites was measured objectively by evaluating participants’ decisions for each of their assigned websites in terms of: 1) avoiding to visit the phishing website (i.e., heeding the warning); 2) clicking the link to open the phishing website homepage but avoiding to browse it; 3) correctly identifying it as a legitimate or phishing website; and 4) avoiding transactions with the phishing website. Each participant was scored as a percentage of correct decisions. Accuracy, run-time, outcome severity of wrong decision, and type of threat were manipulated, and the corresponding information was provided to participants in the experiment.

In terms of controls, the items for security habit were adopted from Pavlou and Fygenson [85], and the items for past encounters with phishing websites and familiarity with domain were developed in this study. Online Appendix D contains the definition of constructs and the sources used for their measurement. Online Appendix E reports the instrument.

The construct items, experiment protocol, and experiment instructions were pretested and pilot-tested. We recruited subjects from multiple groups—university students at a Midwestern university, staff, faculty and the community. In order to reach the community, we posted flyers and placed ads on social media (e.g., Craigslist). We also used the word-of-mouth approach to recruit participants from local communities. The recruitment resulted in a total of 865 participants. Online Appendix F reports the participant profiles. Participants’ education ranged from no degree to doctoral degree, with 72 % falling in the “some college/college student” category. This category had the highest percentage of daily Internet use. The age of participants ranged from 18 to above 58 years, with 88 % falling in 18–25 age category. Gender distribution was 62 % male and 38 % female.

The Internet-use data in the U.S. for 2019 shows that 100 % young adults between the age of 18–29 are Internet users[1] and are the highest users of social media.[2] Young adults with some college education or a college degree have the highest rate of Internet use,[3] making this group the most vulnerable to website-phishing attacks. Indeed, research shows that college students frequently fall victim to various online threats including phishing [16]. Additionally, research has found that the results from student samples are consistent with those from the public panel [86]. Thus, we consider our sample to be suitable for testing the model.

7. Analysis and results

Prior to validating our trust calibration (TC) model, we conducted a series of analyses to check our experiment manipulations, trust calibration, and construct validity. First, we performed a series of ANOVAs on the manipulated variables based on participants’ responses to our manipulation check questions. The questions asked participants to validate their assigned level of detector accuracy, run-time, and outcome severity in the experiment.[4] As shown in Online Appendix G, all the ANOVA tests were significant, indicating that our manipulations were successful.

Second, we assessed whether participants calibrated their trust during the experiment. Specifically, we collected data on participants’ disagreement with the detector’s recommendations about whether a website was safe to visit. We examined how participants adjusted their disagreement during the 10 trials in two groups: high accuracy (i.e., the detector with 90 % accuracy) and low accuracy (i.e., 60 % accuracy) groups. Here, disagreement refers to situations where the participant deemed the website to be legitimate and the detector considered it to be a phishing, or vice versa. Fig. 5 shows the range as well as the mean percentage of disagreement rates (y-axis) for each of the 10 websites encountered by participants in the 10 repeated trails. The mean is presented as a dot, and the range is shown as a vertical bar. The x-axis depicts the order in which the participant made his/her final decision (e.g., 1=the first trial). The chart on the right shows percentage disagreement rates for participants in the high accuracy group. The chart on the left shows disagreement rates for those in the low accuracy group.

In the low accuracy group, the range of disagreement rate was 23 %–33 % with a mean of 28 % in the first trial. The disagreement rate steadily increased over the trials. By the 10th trial, the range was 33 %–44 % with a mean of 38 %. Hence, the results indicate that participants’ trust in the detector was calibrated over the 10 trials.

In the high accuracy group, the range of disagreement rate started at 15 %–24 % with a mean of 20 % at the first trial, fluctuated over the trials, and at the end remained in a similar range—20 %-28 % with a mean of 19 %. Given that the sequence of safe vs. unsafe websites was random and varied for each participant, the presence of warnings about

[1] www.statista.com/statistics/266587/percentage-of-internet-users-by-age-groups-in-the-us/ accessed 5/31/2020.

[2] https://www.statista.com/statistics/471370/us-adults-who-use-social-networks-age/ accessed 5/31/2020.

[3] https://www.pewresearch.org/internet/fact-sheet/internet-broadband/ accessed 5/31/2020.

[4] Type of threat did not have a validation question as it varied in each task.

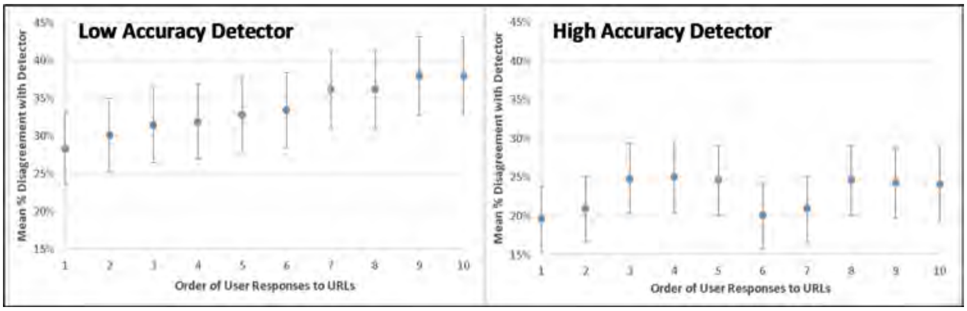


Fig. 5. User Percentage Disagreement with Detector by Trial.

unsafe websites varied for each participant. Such variations could cause small random fluctuations. These results shed light on how users' trust was calibrated during the experiment, albeit at an aggregate level.

Third, we assessed construct reliability, as reported in Table 2. All alpha values were above the threshold of .70 [87], composite factor reliability (CFR) values were greater than the cutoff value of .70 [88], and the average variance extracted (AVE) values were above the threshold of .50 [88], providing support for construct reliability. In addition, we conducted exploratory factor analyses (EFAs) to assess convergent and discriminant validities of our experimental constructs, including controls.

As reported in Online Appendix H, all items loaded on their respective constructs as expected, with all loadings greater than .85 and no cross loadings greater than .40. These results support the convergent and discriminant validity of our constructs [30].

We also compared the square root of the AVE for each construct with its correlations with all other constructs and found that each construct's square root of AVE was greater than the correlation values with the other constructs, as reported in Table 3. The results lend further credence to the discriminant validity of our constructs.

To counter the possibility of common method variance (CMV), we collected perceptual data both before and after the experiment. We also developed the instrument items using semantic differential scales. Moreover, we incorporated an objective variable in the model—ability to detect phishing websites—which was likely to further reduce the threat of CMV. It is worth noting that, in the EFA analysis, no single factor emerged as dominant. Finally, we used a marker variable in our instrument to purify our data prior to analysis [80,89]. The purified data was used in our model estimation to partial-out any potential CMV [80, 89]. With all these remedies, we believe that CMV did not pose a major threat to this study.

We used the structural equations modeling (SEM) as the estimation method due to the fact that the TC model has multiple latent variables and accounts for a number of simultaneous equations involving antecedents, consequents, and control variables of calibrated trust. We used SEM Group analysis in MPlus software. This SEM Group method allows for further simultaneity by estimating the two domains (online pharmacy and online bank) as two distinct groups in the same estimation process. This controls for any dependency that may exist across equations and groups. The estimation method for both the measurement model and the TC model was the mean-adjusted maximum likelihood (MLM) method in MPlus. MLM adjusts the estimations for non-normality in the data. Online Appendix I reports the factor loadings in the measurement model. All factor loadings were above .80 with significant *t*-statistics and high R^2 values, providing further support for the discriminant and convergent validity of the constructs. Fit indices of the measurement model are reported in the second column of Table 4.

Table 2
Construct Reliability Checks.

Constructs	Pharmacy			Bank		
	Cronbach's α	CFR	AVE	Cronbach's α	CFR	AVE
Calibrated trust in the detector	0.97	0.97	0.92	0.97	0.97	0.92
Reliance on detector	0.94	0.96	0.86	0.94	0.96	0.87
Intention to use in the future	0.98	0.98	0.95	0.98	0.98	0.94
Past encounters with phishing site websites	0.88	0.89	0.74	0.91	0.91	0.78
Security habit	0.95	0.95	0.87	0.96	0.96	0.88

Table 3
Construct Correlations and Comparison with Square Root of AVEs.

Constructs (Pharmacy)	1	2	3	4	5
1. Past encounters with phishing websites	**0.86**[a]				
2. Security habit	−0.01	**0.93**			
3. Calibrated trust in the detector	−0.01	−0.03	**0.93**		
4. Reliance on detector	−0.02	−0.06	0.51	**0.96**	
5. Intention to use in the future	−0.02	−0.04	0.45	0.74	**0.98**
Constructs (Bank)	**1**	**2**	**3**	**4**	**5**
1. Past encounters with phishing websites	**0.88**				
2. Security habit	0.03	**0.94**			
3. Calibrated trust in the detector	0.01	0.06	**0.93**		
4. Reliance on detector	0.03	0.12	0.48	**0.96**	
5. Intention to use in the future	0.02	0.09	0.46	0.75	**0.97**

[a] The square root values of the AVEs are reported in boldface on the diagonal.

All values fell within the desired thresholds and supported the model fit. Moreover, the two groups contributed equally to the estimated chi-square, indicating equal fit. We also tested and successfully confirmed measurement invariance between the two groups [90].

The estimation of the TC model also had satisfactory fit indices, as shown in the third column of Table 4. The two groups had approximately equal contributions to the chi-square value, which shows that the model fit was equally satisfactory for both pharmacy and bank domains.

Table 4
Fit Indices for the Measurement Model and TC Model Estimations.

Fit Index	Measurement Model	TC Model	Threshold[a]
Normed $\chi 2$	1.19	1.68	<3
CFI (Comparative Fit Index)	0.997	0.986	>0.90
TLI (Tucker-Lewis Index)	0.997	0.984	>0.90
RMSEA (Root Mean Square Error of Approximation)	0.021	0.040	<0.06
SRMR (Standardized Root Mean Square Residual)	0.021	0.051	<0.10

[a] Gefen et al. [105].

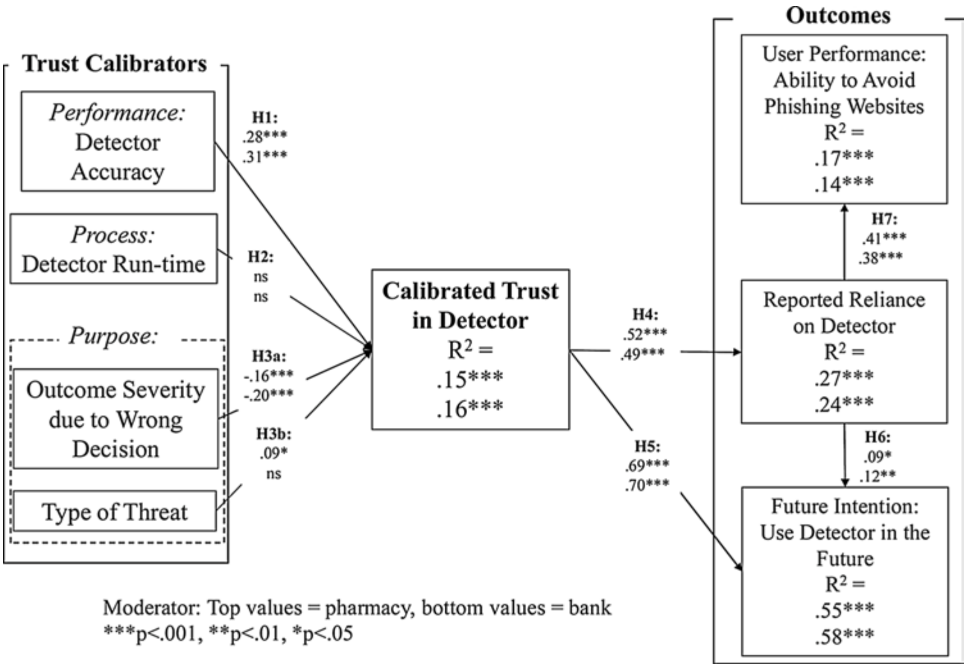

Trust Calibrators
Performance: Detector Accuracy
Process: Detector Run-time
Purpose:
Outcome Severity due to Wrong Decision
Type of Threat
H1: .28*** .31***
H2: ns ns
H3a: -.16*** -.20***
H3b: .09* ns
Calibrated Trust in Detector R2 = .15*** .16***
H4: .52*** .49***
H5: .69*** .70***
Outcomes
User Performance: Ability to Avoid Phishing Websites R2 = .17*** .14***
H7: .41*** .38***
Reported Reliance on Detector R2 = .27*** .24***
H6: .09* .12**
Future Intention: Use Detector in the Future R2 = .55*** .58***
Moderator: Top values = pharmacy, bottom values = bank
***p<.001, **p<.01, *p<.05

Fig. 6. Estimated Trust Calibration Model.

Fig. 6 shows the estimated TC model, reporting path coefficients, p-values, and R^2 values. The top values in Fig. 6 are for the online pharmacy domain, and the bottom values are for the online bank domain. All R^2 values of the endogenous variables in the model were statistically significant in both domains, showing that the TC model had reasonable explanatory power. The TC model estimation supported our conceptual model: all hypotheses were statistically significant in both domains, with the exception of H2 in both domains and H3b in the bank domain.

Hypothesis H1, the influence of the trust calibrator—detector accuracy—on trust in the detector, was supported in both online pharmacy and online bank domains, with path coefficients of .28 in the online pharmacy and .31 in the online bank domain. Surprisingly, hypothesis H2 was not supported as run-time speed had no significant calibrating impact on trust in the detector.

Hypothesis H3a was supported in both domains such that more severe outcomes reduced calibrated trust. Hypothesis H3b was partially supported. In the online pharmacy domain, users showed significant differences in calibrated trust based on type of threat.

With respect to H4 and H5, our findings demonstrated that calibrated trust was positively associated with users' reliance on the detector's advice (H4) and intention of future use (H5), with high path coefficient values across both domains. H4 had path coefficients of .52 (pharmacy) and .49 (bank), and H5 had path coefficients of .69 (pharmacy) and .70 (bank).

Hypotheses H6 and H7 postulated the influence of reliance on future use intention and on users' ability to detect phishing websites. Both hypotheses were supported across both domains, with path coefficients of .09 (pharmacy) and .12 (bank) in H6 and path coefficients of .41 (pharmacy) and .38 (bank) in H7.

Of the control variables, security habit had a significant positive impact on trust in the detector in the online bank domain only, with a path coefficient of .13 and $p < 0.01$, indicating that security habit has a lock-in effect—users habitually act based on prior knowledge. One possible reason that security habit was not significant in the online pharmacy domain is that our participants were relatively young and likely healthy, meaning they had less experience with online pharmacies. Familiarity with domain had significant positive effects on calibrated trust in the detector in the online pharmacy domain—with a path coefficient of .14 and $p < 0.01$. The results showed that those who were more familiar with the online pharmacy domain trusted more in the detector. Finally, gender was significantly associated with trust in the detector in the online pharmacy domain—with a path coefficient of .15 and $p < 0.01$—in that female participants showed higher trust. Age, education, and past encounters with phishing websites showed no significant effects on calibrated trust.

8. Discussion

We developed the trust calibration model for detectors (the TC model) by using the automation trust and reliance (ATR) framework as a kernel theory [33]. The focus was calibrated trust and its antecedents (trust calibrators H1-H3) and consequents (reliance, observed user performance, and future intention to use H4-H7).

8.1. Trust calibrators

Regarding the first set of hypotheses, we found that Hypothesis H1 was fully supported. The results showed that tool accuracy as a performance-based trust calibrator (i.e. informing users of it) was significant across both online pharmacy and online bank domains, with users showing greater trust when the tool accuracy was 90 % than when it was 60 %. This finding underscores the point that users need to be informed about the accuracy of automated tools to prevent their misjudgment on the capability of tools and consequent improper trust and improper reliance [51]. Using accuracy as a trust calibrator is particularly important for automated security IT that detects and eliminates security threats, such as anti-phishing software, as users have consistently shown a predilection to ignore such tools even when the tool's accuracy is high [1].

According to Hypothesis H2, we expected that tool run-time—a process-based trust calibrator—would also influence the level of trust that users placed in the tool. This hypothesis was not supported in either of the two domains. There are several possible explanations for this finding. It is possible that the difference between the one-second and four-second delay was not sufficient for users to consider run-time to be a concern. Another explanation is that users may follow different rationales to interpret run-time depending on how they understand the complexity of the detection task carried out by the tool. For example, a slow run-time may be interpreted as 1) the tool's algorithm being complex and needing a longer time to process or 2) the tool being poorly developed and therefore inefficient and slow. The third explanation is that when users have security in mind, they may be willing to tolerate a few seconds of delay. Lastly, it is also possible that users attribute the delay to other factors, such as a slow Internet or web server, instead of to the detector.

Hypothesis H3a was fully supported across both domains. Our results confirmed outcome severity as a purpose-based trust calibrator—more severe outcomes are associated with greater decreased trust in the tool than less severe outcomes. This finding is important as users often ignore tool warnings despite the fact that the cost of a single security incident may be exceedingly high [1]. In an effort to create proper tool reliance, designers of automated security IT artifacts need to spend additional effort making users aware of the magnitude of risk associated with ignoring tool warnings. The design of security warnings should draw broadly from the literature across domains, including findings on how best to display viscerally aversive warnings [91]. With such design, we could improve users' understanding of their decision outcome in the context of automated security IT.

In accordance with Hypothesis H3b, we found that users in the online pharmacy domain showed greater trust in the detector for concocted websites than for spoof websites, while users in the online banking domain showed no significant difference in trust based on type of threat. This finding suggests that users may tend to trust the detector more in detecting concocted than spoof phishing websites within certain domains. One possible explanation for this is that in unfamiliar domains, users do not have a well-known legitimate online entity they can reference when detecting a concocted site. Consequently, they may rely more on the tool and less on themselves to assess the credibility of the novel website as they have less existing information to draw upon.

Comparing the standardized path coefficients in the estimated TC model, we found the effects of our antecedent trust calibrators on trust are different based on the relevant path loadings (see Fig. 6). Detector accuracy and outcome severity have higher path coefficients and are more effective trust calibrators than detector run-time and threat type. This finding implies that when users assess the key characteristics of automated security IT to form trust, their assessment is based more on the performance (accuracy) and purpose (outcome severity due to wrong decision) calibrators than the process calibrators. Designers of automated security IT artifacts would benefit from the design of warning displays that emphasize the most effective trust calibrators, which in our case are performance and purpose calibraators. Designers may also benefit from designs that provide users with information on the most effective trust calibrators to build their proper trust.

We also examined the effect size of trust calibrators on calibrated trust for those significant paths and found that the effect size (f^2) values ranged from 0.015 to 0.115 (see Online Appendix J). A 30-year review on effect size [92] shows that the median observed effect size (f^2) is .002. Thus, we argue that our trust calibrators are effective in calibrating trust with the above median effect size.

Finally, we conducted a post hoc analysis on the interaction effect of trust calibrators on calibrated trust and found no significant interaction effect.

8.2. Outcomes

With regard to the latter four hypotheses concerning the effects of calibrated trust on user performance outcomes, all found support within the TC model across both online pharmacy and online banking domains. First, the findings from H4 and H5 confirmed the strong positive association between trust in the detector and reliance on it and between trust in the detector and future intention to use it. Along with the findings from H1-H3, the significance of hypotheses H4 and H5 underscores the utility and value of the TC model: with proper antecedent trust calibrators in place, it is possible to change users' security behaviors and ensure their appropriate reliance on the tool's advice. Such changes could be accomplished through the paths from trust calibrators→calibrated trust in detector→ reliance on and use of detector. These findings highlight that trust in the detector is a central conduit that connects trust calibrators to desirable user behavior and improved tool performance. The findings also imply that designers of automated security IT need to be transparent about their tools and manage users' mental models to foster their proper understanding of security IT characteristics. Otherwise, users tend to be self-reliant even when the automated tool outperforms them [1].

We conducted a post hoc analysis to validate the mediating role of calibrated trust in bridging trust calibrators and desirable outcomes from the detector. We used the bootstrapping mediation test in Mplus for the analysis [93]. As shown in Online Appendix K, all mediating effects were significant when the path from the trust calibrator to calibrated trust was significant. The findings further confirm the effects of the trust calibrators and the central role of calibrated trust.

Phishing website detection tools are known to suffer from disuse and suboptimal performance even when the tool's accuracy is high [1,13,14,22]. Thus, to change users' security behaviors and increase their reliance on the advice of tools, it is critical that tool designers effectively improve trust calibrators and inform users about them. In this respect, our study answers a call for research on "active exploration for trusting" (AET), a methodology that promotes frequent trust calibration and enables trust calibration from an ante hoc perspective [32]. We may need more empirical and analytical methods to identify users' inappropriate trust and ongoing misuse or disuse behaviors in a timely manner so that we can promptly initiate the trust calibration process.

Reliance on the tool has two important positive consequences: future intention to use (H6) and an observed increase in user performance (H7). Comparing the standardized path coefficients in Fig. 6, we found that reliance on the detector leads to higher user performance in avoiding phishing websites compared to a weaker path coefficient of future intention to use. This finding is interesting, as it reveals the tenuous nature of human-to-automation interactions. While human-to-human interactions with positive outcomes may lead to strong loyalty [94], human-to-automation interactions do not benefit from the emotional bonds found in interpersonal relationships. When it comes to users' reliance on automated tools, today's acceptance of advice does not necessarily translate into future use. Any misjudgment developed during usage could break prior established trust. This finding is a microcosm of a larger problem with enterprise security: while security managers are constantly looking to upgrade existing security IT artifacts and add new ones, they also have to continually increase employees' security awareness and motivation to use security tools and comply with security policies [95]. According to a Gartner survey [96], enterprise security expenditures increased an estimated 8% each year since 2016 due to the persistent threat landscape and more high-profile cyberattacks. Our findings suggest that increasing expenditures with the goal of having more advanced automated security systems including detection systems may be just one pillar required to achieve strong security. Routinely educating employees and calibrating their trust in such systems for proper use and reliance may be a second key pillar.

Another positive consequence of reliance on the detector is users' improved ability to detect phishing websites, per H7. People generally perform poorly when it comes to detecting deception, and individuals are particularly ill-equipped to detect phishing websites [1,15]. Our findings regarding H7, combined with the findings from H1-H4, provide holistic support and value for the TC model: calibrated trust based on carefully selected trust calibrators can improve users' detection capabilities.

Specifically, we have shown an important pathway that could lead to increased tool performance: carefully identified and employed trust calibrators→calibrated trust→proper reliance on tool advice→improved

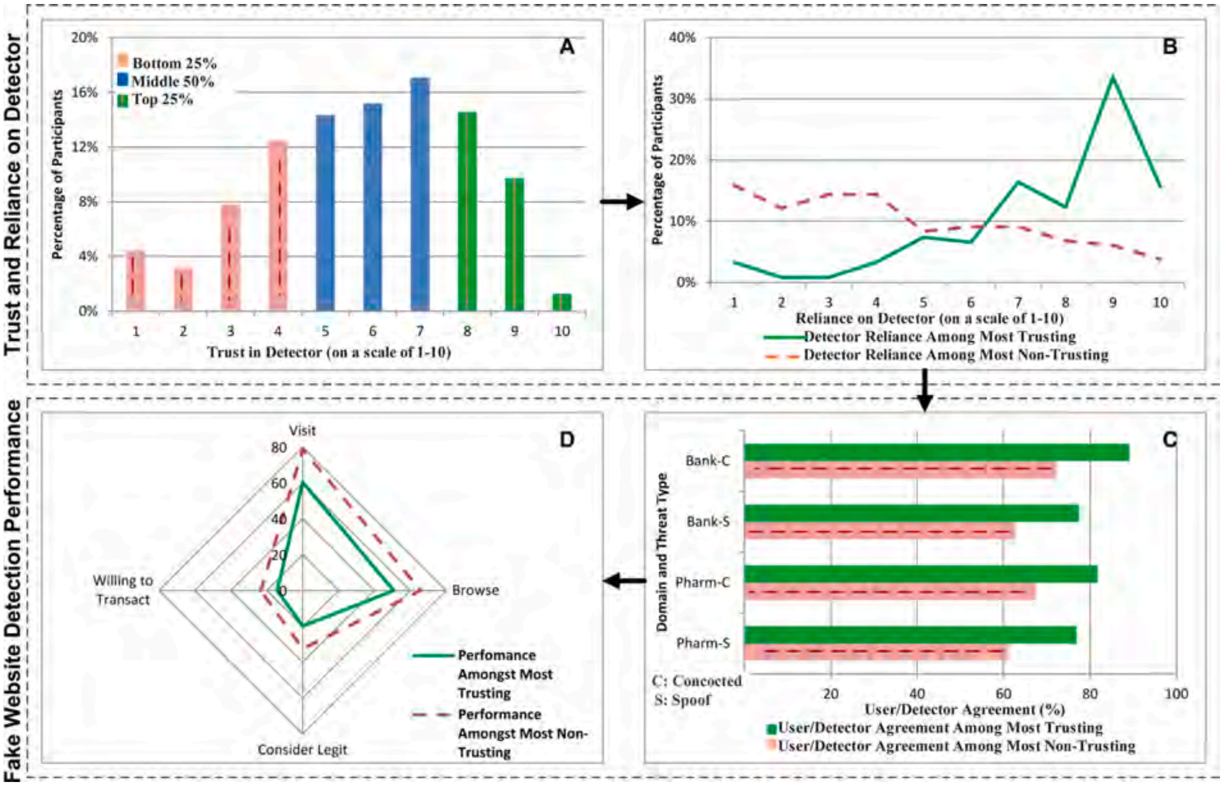


Fig. 7. Appropriate Trust – Impact of Trust in Detector on Reliance and Performance.

users' ability to avoid phishing websites.

Fig. 7 further illustrates this pathway for participants using the 90 % accurate detector (averaged across the three antecedent trust calibrators of accuracy, severity of decision outcome, and run-time). We divided participants into quartiles based on their calibrated trust (measured on a scale of 1–10). Panel A in Fig. 7 shows the histogram of calibrated trust with the top quartile (top 25 %) in solid-line columns (green color) and the bottom quartile (bottom 25 %) in dash-line columns (peach color). Panel B compares the tool reliance of the top and bottom trusting quartiles. Our calculations showed that 78 % of the top-quartile users (solid line) reported high use of the tool as compared to only 26 % of the bottom-quartile users (dash-line). Here, we define 'high use' as an average of 7 or greater on a 1–10 scale. Panel C shows the distribution of participants' agreement with detector for the top (solid line) and bottom (dash-line) quartiles. Likewise, we found the most trusting users (top quartile) were generally 20%–25% more likely to agree with and heed the tool's recommendations regarding spoof and concocted websites associated with either domain based on objective performance data.

Panel D shows participants' performance for the top and bottom quartiles. Our computation showed that compared to the least trusting users (bottom quartile), the most trusting users (top quartile) were 23 % less likely to visit phishing websites, 22 % less likely to browse multiple pages on them, 38 % less likely to consider phishing websites legitimate, and 39 % less willing to transact with phishing websites. Collectively, Fig. 7 panels demonstrate our key points when applied to high performing tools (e.g., 90 % accuracy): users' trust in detector tools should be calibrated, users' calibrated trust aligns with the tools' performance, and calibrated trust leads to increased users' reliance on the tool and better ability to avoid threats.

9. Theoretical and practical implications

This study has significant theoretical and practical implications as discussed below.

9.1. Theoretical implications

This study makes several contributions to IS research. First, this study contributes to IS literature on trust in security IT. By identifying phishing-website security tools as a type of automated security IT with user choice, this study introduces calibrated trust as a type of trust that needs to be calibrated by the trust calibrators to align with the capability of detection tools. Prior research in other fields such as human factors has long recognized the significance of this concept in automation, including in new automated systems (e.g., adaptive cruise control) [38, 39,45]. However, IS research has been silent about trust calibration in security IT. Our study is a significant addition to IS research in trust and opens an avenue for the theory-based research in this area.

Second, this study argues for a theory-guided trust calibration for automated security IT, and applies the key theme of automation trust and reliance (ATR) framework into the context of phishing-website detection tools as exemplars of automated security IT. To the best of our knowledge, the ATR framework has not been applied to the IS field, especially to the automated security IT research. The contextualization of this theory to the study of automated security IT provides a new theoretical foundation for future research in this area [75,97,98]. Through the contextualization, we propose a theoretical model for trust calibration for phishing-website detection tools (referred to as the TC model). The TC model contributes to theory in several ways. It identifies salient antecedent trust calibrators that are necessary for users to form an appropriate degree of trust in an automated security detection tool. It provides a theoretical basis for the need to inform users about salient trust calibrators to promote an appropriate level of users' trust in detection tools. It provides insight into consequent performance outcomes of properly calibrated trust, including tool reliance, future usage intentions, and improved detection performance. It builds a theoretical foundation for trust calibration in the context of automated security IT. More importantly, it presents a key pathway from identifying and employing trust calibrators of a security tool, to calibrating proper user trust in the tool, to achieving desirable performance outcomes. Thus, the contextualization of ATR and the resulting TC model are significant theoretical contributions to trust research, especially trust research in security IT.

Third, the empirical validation of the TC model also contributes to IS research. More importantly, the study provides empirical evidence that we can calibrate trust in the context of automated security IT that allows users to make final decisions. With this evidence in hand, this study highlights "calibratability" of trust in security IT, an area that has not received much attention from IS researchers.

Lastly, the TC model can be used to guide the investigation of trust and trust calibration for other security tools and systems. More specifically, the TC model can serve as a theoretical model to guide the selection of proper calibrators for trust calibration in other security IT artifacts and validate their effectiveness. In addition, the TC model may be applied to other automated detection tools and systems for trust calibration, such as deception detection systems [99] and automated interviewing systems [100]. Detection accuracy of such tools and systems is well below 100 %, and thus a proper level of trust needs to be established for increased human-automation performance.

9.2. Practical implications

This study addresses a call for research in the relationships between trust and IT in general and between trust and automated security IT in particular [26,101]. As shown in Fig. 7, while using the tool with the same accuracy, users demonstrated significant differences in trust and detection performance. Thus, when designing automated security IT artifacts, designers must inform users of trust calibrators, particularly the tools' detection accuracy rate. An implementation method, we suggest to deliver trust calibrator information to users through warning messages (see Panel A in Fig. 4). Warning science suggests that an effective warning text may consist of four types of message information: a signal word, description of the threat, potential negative consequences, and instructions on how to avoid the hazard [91]. The information on trust calibrators can be part of the instruction information of warning messges devlivered by the detector. Designers can carefully design warnings and their displays to create calibration effects and keep users in the feedback loop when using the detector. Vendors of detection tools with a high detection rate can gain a business edge by publicizing their high detection rates via finding effective ways to communicate them to their customers in use of the tool.

The results of this study also have important implications for IT managers, Chief Information Officers (CIOs), and Chief Security Officers (CSOs) tasked with enterprise security. Based on findings pertaining to the TC model, organizations must consider together the following two avenues of their security practice:

1 *Adopting accurate security systems/tools with effective trust calibration features whenever possible.* In the context of phishing website detection, benchmarking studies show that state-of-the-art tools are approximately 95 % accurate, including proprietary and enterprise-grade tools [1,102,103]. Relatively low accuracy can be found in some automated systems dealing with complex decisions and tasks [104]. Moreover, even with a highly accurate but "imperfect" automated system, users still demonstrate low performance due to unjustified trust and reliance [17]. Thus, when investing in high performance security systems/tools, organizations need to understand the subjective nature of trust in automated security IT and consider adopting state-of-the art security systems/tools with built-in design features that facilitate trust calibration as much as possible. For example, it may be worth preferring tools that provide "visible" feedback to users when necessary (e.g., when detecting a high click-through rate and low reliance). Ultimately, the key is to allow the user to establish an appropriate level of trust in the tool that is consistent with its capabilities.

2 *Allocating resources for education and training programs about security and trust calibration of security systems/tools.* Academics and practitioners both agree that removing users entirely from the security loop sometimes is impractical, particularly in the context of phishing website detection [11]. Based on our findings regarding the TC model, organizations need to appreciate the importance of trust calibration and allocate resources for security training and education programs that highlight the performance, process, and purpose of security systems/tools to foster employees' better understanding of automated security tools' capabilities—a fundamental prerequisite for establishing an appropriate level of trust in automated tools. Security education and training programs should not only enhance knowledge pertaining to security threats such as phishing websites, but endorse a solution- and tool-centric training paradigm in which trust calibration is purposefully incorporated. The TC model can serve as an evaluative model to assess the effectiveness of such training.

Our work also has implications for experts and individual users. Our work has highlighted the importance of trust calibrators for increased use and protection against phishing websites. Security experts need to increase their focus on detailed reviews of security detections tools in terms of trust calibrators. In order to increase their protection performance, individuals should insist on having the detailed information about trust calibrators and give preference to vendors who provide such information.

10. Limitations and future research directions

There are limitations in this study. Although we conducted our experiment in two distinct domains—online pharmacies and online banks—care should be taken when generalizing our results to other domains. In addition, our participants interacted with a complex and custom-written detection tool (by the necessity of controlled experimental design) instead of interacting with a well-known existing program or plug-in. Further care should be taken to consider possible influences of tool brand names and participants' varying experiences and prior knowledge of such tools. Moreover, this study was conducted in the U.S., and our sample consisted mainly of young participants. Research should replicate our work with data from other countries and older populations. Moreover, calibrated trust can fluctuate under different conditions, and can change along with long-term interactions with the system [32,33,37,38]. It would be of interest to examine how calibrated trust fluctuates as a result of experiencing loss due to attacks and other events while using the tools. Future research needs to develop measures specific to calibrated trust in the context of automated security tools. Furthermore, for the sake of realism, we used a progress bar to visualize run-time, representing the tool's process. More work is needed to identify additional proxies for the inner processes of detection tools, including alternative methods of informing users about the detector's run-time, such as a progress bar with text information, animation, or other visualization methods.

Our research opens several new research avenues. First, the TC model demonstrates the importance of identifying proper trust calibrators. As shown in our results, not all calibrators exhibited significant calibrating effects. In our case, the tool's performance-based feature of accuracy and its purpose-based feature of decision outcome severity are more influential calibrators than the other two calibrators. Thus, more research may be needed to understand the difference in effect size of trust calibrators by addressing the following research question: Do all trust calibrators exhibit the same (or similar) calibrating effects on automated security IT artifacts that perform different tasks or have different levels in user control? Answering this question would help further strengthen our understanding within the TC model framework.

Second, there is no endpoint to trust calibration, and trust evolves over time [37,38]. Researchers need to investigate how trust in automated security IT with user choice changes in response to personal and social events and experiences. Additionally, once a disuse behavior or overuse behavior has already occurred, it is often too late to change the damaging effects of the behavior via trust calibration [68]. Therefore, finding ways to identify a user's misplaced trust and the best time to calibrate/recalibrate it to build and maintain an appropriate level of trust remains a challenging research question that warrants an answer [32,37]. Indeed, finding the optimal level of trust is also a research challenge [32]. Moreover, finding an effective way to convey the information about trust is another interesting topic for future research.

Further, finding an effective way to convey the information of trust calibrators of a security tool to users (e.g., through security warnings) is also an important topic. Both warning sciences and Lee and See [33] suggest that where, how, and when to display such information could influence the effect of trust calibration. The design of the interface and placement of the trust calibrators constitute important areas for further research.

Another direction for future research is to evaluate the TC model in alternate domains such as social media, which is increasingly being targeted by phishing attacks and has important strategic implications for organizations that use it for internal communication (e.g., Slack) or customer engagement and support (e.g., Twitter). Moreover, this study focuses on individual users without exploring trust calibrators related to other external contexts—such as organizational and cultural contexts. Future research may also consider investigating the trust calibration effects on security IT in such contexts.

Finally, it would also be beneficial to explore the influence of different IT platforms on security behavior and trust calibration of security IT, such as mobile computing, cloud computing, and other platforms. In all these cases, the TC model may provide a clear path for conducting future research along these avenues.

Acknowledgements

This work was funded in part by a grant from the U.S. National Science Foundation under Grant No. CNS-1049497. We would also like to thank our partners at McAfee Security for their invaluable feedback on the constructs, experiment design, survey instrument, and data analysis.

Supplementary Material: Online Appendixes

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.im.2020.103394.

References

[1] A. Abbasi, F. Zahedi, D. Zeng, Y. Chen, H.C. Chen, J.F.E. Nunamaker, Enhancing predictive analytics for anti-phishing by exploiting website genre information, J. Manag. Inf. Syst. 31 (4) (2015) 109–157.

[2] E.D. Frauenstein, S. Flowerday, Susceptibility to phishing on social network sites: a personality information processing model, Comput. Secur. (2020), 101862.

[3] F.M. Zahedi, A. Abbasi, Y. Chen, Fake-website detection tools: identifying elements that promote individuals' use and enhance their performance, J. Assoc. Inf. Syst. 16 (6) (2015) 448–484.

[4] FBI, Business Email Compromise the $26 Billion Scam, 2019. Available at https://www.ic3.gov/media/2019/190910.aspx. (Accessed 18 Feburary 2020).

[5] RiskIQ, The Evil Internet Minute, 2019. Available at https://www.riskiq.com/infographic/evil-internet-minute-2019/. (Accessed 10 May 2020).

[6] G. Aaron, R. Rasmussen, Global Phishing Survey: Trends and Domain Name Use in 2016, APWG (2017). June 26. Available at https://docs.apwg.org/reports/APWG_Global_Phishing_Report_2015-2016.pdf. (Accessed 18 August 2018).

[7] A. Abbasi, Z. Zhang, D. Zimbra, H. Chen, J.F.D. Nunamaker, Etecting fake websites: the contribution of statistical learning theory, MIS Q. 34 (3) (2010) 435–461.

[8] Z. Whittaker, Google Let Scammers Post a Perfectly Spoofed Amazon Ad in its Search Results, ZDNet, 2017. Available at http://www.zdnet.com/article/malicious-google-ad-pointed-millions-to-fake-windows-support-scam/. (Accessed 8 March 2019).

[9] A. Abbasi, F. Zahedi, Y. Chen, Impact of anti-phishing tool performance on attack success rates, Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on: IEEE (2012) 12–17.

[10] H. Cavusoglu, S. Raghunathan, Configuration of detection software: a comparison of decision and game theory approaches, Decis. Anal. 1 (3) (2004) 131–148.

[11] L.F. Cranor, A framework for reasoning about the human in the loop, in: The Conference on Usability, Psychology, and Security, Berkeley, CA, 2008.

[12] C. Bravo-Lillo, S. Komanduri, L.F. Cranor, R.W. Reeder, M. Sleeper, J. Downs, S. Schechter, Your attention please: designing security-decision uis to make genuine risks harder to ignore, in: Proceedings of the Ninth Symposium on Usable Privacy and Security, ACM, 2013.

[13] R.W. Reeder, A.P. Felt, S. Consolvo, N. Malkin, C. Thompson, S. Egelman, An experience sampling study of user reactions to browser warnings in the field, Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (2018) 1–13.

[14] Symantec, The Norton Cybercrime Report: The Human Impact, 2010. Available at https://www.symantec.com/content/en/us/home_homeoffice/media/pdf/cybercrime_report/Norton_USA-Human%20Impact-A4_Aug4-2.pdf. (Accessed 8 December 2018).

[15] R. Dhamija, J.D. Tygar, M. Hearst, Why phishing works, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Montreal, Canada: ACM, 2006, pp. 581–590.

[16] S. Goel, K. Williams, E. Dincelli, Got phished? Internet security and human vulnerability, J. Assoc. Inf. Syst. 18 (1) (2017) 22–44.

[17] D. Akhawe, A.P. Felt, Alice in warningland: a large-scale field study of browser security warning effectiveness, USENIX Security Symposium (2013).

[18] D. Cyr, Modeling web site design across cultures: relationships to trust, satisfaction, and e-loyalty, J. Manag. Inf. Syst. 24 (4) (2008) 47–72.

[19] D. Gefen, E. Karahanna, D.W. Straub, Trust and TAM in online shopping: an integrated model, MIS Q. 27 (1) (2003) 51–90.

[20] N.K. Lankton, D.H. McKnight, J. Tripp, Technology, humanness, and trust: rethinking trust in technology, J. Assoc. Inf. Syst. 16 (10) (2015) 880–918.

[21] J. Song, F. Zahedi, Dynamics of trust revision: using health infomediaries, J. Manag. Inf. Syst. 24 (4) (2008) 225–248.

[22] Y. Chen, F.M. Zahedi, A. Abbasi, Interface design elements for anti-phishing systems, in: International Conference on Design Science Research in Information Systems, Springer, 2011, pp. 253–265.

[23] B.B. Anderson, A. Vance, C.B. Kirwan, D. Eargle, J.L. Jenkins, How users perceive and respond to security messages: a neurois research agenda and empirical study, Eur. J. Inf. Syst. 25 (4) (2016) 364–390.

[24] A. Vance, J.L. Jenkins, B.B. Anderson, D.K. Bjornn, C.B. Kirwan, Tuning out security warnings: alongitudinal examination of habituation through fmri, eye tracking, and field experiments, MIS Q. 42 (2) (2018) 355–380.

[25] N.K. Lankton, D.H. McKnight, R.T. Wright, J.B. Thatcher, Using expectation disconfirmation theory and polynomial modeling to understand trust in technology, Inf. Syst. Res. 27 (1) (2016) 197–213.

[26] D.H. McKnight, M. Carter, J.B. Thatcher, P.F. Clay, Trust in a specific technology: an investigation of its components and measures, ACM Trans. Manage. Inf. Syst. (TMIS) 2 (2) (2011) 12.

[27] A. Vance, C. Elie-Dit-Cosaque, D.W. Straub, Examining trust in information technology artifacts: the effects of system quality and culture, J. Manag. Inf. Syst. 24 (4) (2008) 73–100.

[28] W.Q. Wang, I. Benbasat, Attributions of trust in decision support technologies: a study of recommendation agents for e-commerce, J. Manag. Inf. Syst. 24 (4) (2008) 249–273.

[29] T. Dinev, Q. Hu, The centrality of awareness in the formation of user behavioral intention toward protective information technologies, J. Assoc. Inf. Syst. 8 (7) (2007) 386–408.

[30] D.H. McKnight, V. Choudhury, C. Kacmar, Developing and validating trust measures for e-commerce: an integrative typology, Inf. Syst. Res. 13 (3) (2002) 334–359.

[31] L. Li, E. Berki, M. Helenius, S. Ovaska, Towards a contingency approach with whitelist-and blacklist-based anti-phishing applications: what do usability tests indicate? Behav. Inf. Technol. 33 (11) (2014) 1136–1147.

[32] R.R. Hoffman, M. Johnson, J.M. Bradshaw, A. Underbrink, Trust in automation, IEEE Intell. Syst. 28 (1) (2013) 84–88.

[33] J.D. Lee, K.A. See, Trust in automation: designing for appropriate reliance, Hum. Factors 46 (1) (2004) 50–80.

[34] P. Madhavan, D.A. Wiegmann, Similarities and differences between human–human and human–automation trust: an integrative review, Theor. Issues Ergon. Sci. 8 (4) (2007) 277–301.

[35] K.E. Schaefer, J.Y. Chen, J.L. Szalma, P.A. Hancock, A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems, Hum. Factors 58 (3) (2016) 377–400.

[36] R. Parasuraman, T.B. Sheridan, C.D. Wickens, A model for types and levels of human interaction with automation, IEEE Trans. Syst. Man Cybernet.-Part A: Syst. Hum. 30 (3) (2000) 286–297.

[37] K.A. Hoff, M. Bashir, Trust in automation: integrating empirical evidence on factors that influence trust, Hum. Factors 57 (3) (2015) 407–434.

[38] J. Kraus, D. Scholz, D. Stiegemeier, M. Baumann, The more you know: trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency, Hum. Factors (2019), 0018720819853686.

[39] Y. Lu, N. Sarter, Eye tracking: a process-oriented method for inferring trust in automation as a function of priming and system reliability, IEEE Trans. Hum. Syst. 49 (6) (2019) 560–568.

[40] J. Chen, S. Mishler, B. Hu, N. Li, R.W. Proctor, The description-experience gap in the effect of warning reliability on user trust and performance in a phishing-detection context, Int. J. Hum. Stud. 119 (2018) 35–47.

[41] E.G. Glikson, A.W. Woolley, Human trust in artificial intelligence: review of empirical research, Acad. Manag. Ann. 14 (2) (2020).

[42] N. Paravastu, D. Gefen, S.B. Creason, Understanding trust in it artifacts: an evaluation of the impact of trustworthiness and trust on satisfaction with antiviral software, in: ACM SIGMIS Database: the DATABASE for Advances in Information Systems, 45, 2014, pp. 30–50, 4.

[43] R. West, The psychology of security, Commun. ACM 51 (4) (2008) 34–40.

[44] A. Whitten, J.D. Tygar, Why johnny can't encrypt: a usability evaluation of pgp 5.0, USENIX Security Symposium (1999).

[45] K. Stowers, N. Kasdaglis, M.A. Rupp, O.B. Newton, J.Y. Chen, M.J. Barnes, The impact of agent transparency on human performance, IEEE Trans. Hum. Syst. 50 (3) (2020) 245–253.

[46] P. De Vries, C. Midden, D. Bouwhuis, The effects of errors on system trust, self-confidence, and the allocation of control in route planning, Int. J. Hum. Stud. 58 (6) (2003) 719–735.

[47] S.M. Merritt, D. Lee, J.L. Unnerstall, K. Huber, Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task, Hum. Factors 57 (1) (2015) 34–47.

[48] E.T. Chancey, J.P. Bliss, Y. Yamani, H.A. Handley, Trust and the compliance–reliance paradigm: the effects of risk, error bias, and reliability on trust and dependence, Hum. Factors 59 (3) (2017) 333–345.

[49] J.M. McGuirl, N.B. Sarter, Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information, Hum. Factors 48 (4) (2006) 656–665.

[50] P.A. Hancock, D.R. Billings, K.E. Schaefer, J.Y. Chen, E.J. De Visser, R. Parasuraman, A meta-analysis of factors affecting trust in human-robot interaction, Hum. Factors 53 (5) (2011) 517–527.

[51] S.M. Merritt, J.L. Unnerstall, D. Lee, K. Huber, Measuring individual differences in the perfect automation schema, Hum. Factors 57 (5) (2015) 740–753.

[52] M.T. Dzindolet, L.G. Pierce, H.P. Beck, L.A. Dawe, The perceived utility of human and automated aids in a visual detection task, Hum. Factors 44 (1) (2002) 79–94.

[53] P. Madhavan, D.A. Wiegmann, F.C. Lacson, Automation failures on tasks easily performed by operators undermine trust in automated aids, Hum. Factors 48 (2) (2006) 241–256.

[54] G. Xiang, J. Hong, C.P. Rose, L. Cranor, Cantina+: a feature-rich machine learning framework for detecting phishing web sites, ACM Trans. Inf. Syst. Security (TISSEC) 14 (2) (2011) 21.

[55] A. Bouch, A. Kuchinsky, N. Bhatti, Quality is in the eye of the beholder: meeting users' requirements for internet quality of service, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, The Hague, Netherlands ACM, 2000, pp. 297–304.

[56] J.L. Jenkins, B.B. Anderson, A. Vance, C.B. Kirwan, D. Eargle, More harm than good? How messages that interrupt can make us vulnerable, Inf. Syst. Res. 27 (4) (2016) 880–896.

[57] J. Hohenstein, H. Khan, K. Canfield, S. Tung, R. Perez Cano, Shorter wait times: the effects of various loading screens on perceived performance, in: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, ACM, 2016, pp. 3084–3090.

[58] S.E. McBride, W.A. Rogers, A.D. Fisk, Understanding human management of automation errors, Theor. Issues Ergon. Sci. 15 (6) (2014) 545–577.

[59] V.L. Pop, A. Shrewsbury, F.T. Durso, Individual differences in the calibration of trust in automation, Hum. Factors 57 (4) (2015) 545–556.

[60] J. Koepke, S. Kaza, A. Abbasi, Exploratory experiments to identify fake websites by using features from the network stack, Intelligence and Security Informatics (ISI), IEEE International Conference on: IEEE (2012) 126–128.

[61] M.T. Khasawneh, S.R. Bowling, X. Jiang, A.K. Gramopadhye, B.J. Melloy, Effect of error severity on human trust in hybrid systems, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE Publications Sage CA: Los Angeles, CA, 2004, pp. 439–443.

[62] P.P. Duez, M.J. Zuliani, G.A. Jamieson, Trust by design: information requirements for appropriate trust in automation, in: Proceedings of the 2006 Conference of the Center for Advanced Studies on Collaborative Research, IBM Corp., 2006.

[63] S. Grazioli, S.L. Jarvenpaa, Consumer and business deception on the internet: content analysis of documentary evidence, Int. J. Electron. Commer. 7 (4) (2003) 93–118.

[64] S.M. Merritt, Affective processes in human–automation interactions, Hum. Factors 53 (4) (2011) 356–370.

[65] A.C. Elkins, N.E. Dunbar, B. Adame, J.F. Nunamaker, Are users threatened by credibility assessment systems? J. Manag. Inf. Syst. 29 (4) (2013) 249–261.

[66] M.L. Jensen, P.B. Lowry, J.L. Jenkins, Effects of automated and participative decision support in computer-aided credibility assessment, J. Manag. Inf. Syst. 28 (1) (2011) 201–233.

[67] K.E. Culley, P. Madhavan, Trust in automation and automation designers: implications for hci and hmi, Comput. Human Behav. 6 (29) (2013) 2208–2210.

[68] K. Drnec, A.R. Marathe, J.R. Lukos, J.S.F. Metcalfe, rom trust in automation to decision neuroscience: applying cognitive neuroscience methods to understand and improve interaction decisions involved in human automation interaction, Front. Hum. Neurosci. 10 (2016) 290.

[69] R. Parasuraman, D.H. Manzey, Complacency and bias in human use of automation: an attentional integration, Hum. Factors 52 (3) (2010) 381–410.

[70] N.K. Lankton, D.H. McKnight, What does it mean to trust facebook?: Examining technology and interpersonal trust beliefs, in: ACM SIGMIS Database: the DATABASE for Advances in Information Systems, 42, 2011, pp. 32–54, 2.

[71] J. Gao, J.D. Lee, Extending the decision field theory to model operators' reliance on automation in supervisory control situations, IEEE Trans. Syst. Man Cybernet.-Part A: Syst. Hum. 36 (5) (2006) 943–959.

[72] S.S. Kim, N.K. Malhotra, S. Narasimhan, Two competing perspectives on automatic use: a theoretical and empirical comparison, Inf. Syst. Res. 16 (4) (2005) 418–432.

[73] M.L. Jensen, P.B. Lowry, J.K. Burgoon, J.F. Nunamaker, Technology dominance in complex decision making: the case of aided credibility assessment, J. Manag. Inf. Syst. 27 (1) (2010) 175–201.

[74] T.N. Jagatic, N.A. Johnson, M. Jakobsson, F. Menczer, Social phishing, Commun. ACM 50 (10) (2007) 94–100.

[75] D.A. Whetten, T. Felin, B.G. King, The practice of theory borrowing in organizational studies: current issues and future directions, J. Manage. 35 (3) (2009) 537–563.

[76] G. Bansal, F. Zahedi, D. Gefen, The moderating influence of privacy concern on the efficacy of privacy assurance mechanisms for building trust: a multiple-context investigation, in: The International Conference on Information Systems (ICIS), Paris, France, 2008.

[77] C. Hellerman, Fda Shuts Down 1,677 Online Pharmacies, CNN, 2013. June 23. Available at https://www.cnn.com/2013/06/27/health/online-pharmacies-closed/index.html. (Accessed 8 May 2018).

[78] A. Greenberg, Pharma's Black Market Boom. Forbes.cOm, 2008. August 26. Available at https://www.forbes.com/2008/08/25/online-pharma-scams-tech-security-cx_ag_0826drugscam.html#140e76d8e709. (Accessed 8 May 2018).

[79] W.W. Chin, N. Johnson, A. Schwarz, A fast form approach to measuring technology acceptance and other constructs, MIS Q. 32 (4) (2008) 687–703.

[80] P.M. Podsakoff, S.B. MacKenzie, J.-Y. Lee, N.P. Podsakoff, Common method biases in behavioral research: a critical review of the literature and recommended remedies, J. Appl. Psychol. 88 (5) (2003) 879.

[81] G. Bansal, F. Zahedi, D. Gefen, The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online, Decis. Support Syst. 49 (2) (2010) 138–150.

[82] M. Madsen, S. Gregor, Measuring human-computer trust, in: 11th Australasian Conference on Information Systems, Sydney, Australia, 2000, pp. 6–8. Citeseer.

[83] F.D. Davis, R.P. Bagozzi, P.R. Warshaw, User acceptance of computer technology: a comparison of two theoretical models, Manage. Sci. 35 (8) (1989) 982–1003.

[84] V. Venkatesh, M.G. Morris, G.B. Davis, F.D. Davis, User acceptance of information technology: toward a unified view, MIS Q. 27 (3) (2003) 425–478.

[85] P.A. Pavlou, M. Fygenson, Understanding and predicting electronic commerce adoption: an extension of the theory of planned behavior, MIS Q. 30 (1) (2006) 115–143.

[86] Z.R. Steelman, B.I. Hammer, M. Limayem, Data collection in the digital age: innovative alternatives to student samples, MIS Q. 38 (2) (2014) 355–378.

[87] J.C. Nunnally, I.H. Bernstein, Psychometric Theory, McGraw Hill, New York, 1978.

[88] A.H. Segars, Assessing the unidimensionality of measurement: a paradigm and illustration within the context of information systems research, Omega 25 (1) (1997) 107–121.

[89] R.P. Bagozzi, Measurement and meaning in information systems and organizational research: methodological and philosophical foundations, MIS Q. 35 (2) (2011) 261–292.

[90] W.J. Doll, A. Hendrickson, X. Deng, Using davis's perceived usefulness and ease-of-use instruments for decision making: a confirmatory and multigroup invariance analysis, Decis. Sci. 29 (4) (1998) 839–869.

[91] A. Black, P. Luna, O. Lund, S. Walker, Information Design: Research and Practice, Routledge, 2017.

[92] H. Aguinis, J.C. Beaty, R.J. Boik, C.A. Pierce, Effect size and power in assessing moderating effects of categorical variables using multiple regression: a 30-year review, J. Appl. Psychol. 90 (1) (2005) 94–107.

[93] L.K. Muthén, B.O. Muthén, Mplus User's Guide, Muthén & Muthén, Los Angeles, CA, 2012.

[94] F.M. Zahedi, G. Bansal, J. Ische, Success factors in cooperative online marketplaces: trust as the social capital and value generator in vendors-exchange relationships, J. Organ. Comput. Electron. Commer. 20 (4) (2010) 295–327.

[95] Y. Chen, K. Ramamurthy, K.W. Wen, Organizations' information security policy compliance: stick or carrot approach? J. Manag. Inf. Syst. 29 (3) (2013) 157–188.

[96] Gartner survey, Gartner Forecasts Worldwide Security Spending Will Reach $96 Billion in 2018, up 8 Percent From 2017, 2017. December 7. Available at https://www.gartner.com/newsroom/id/3836563. (Accessed 8 January 2018).

[97] Y. Chen, F.M. Zahedi, Individuals' internet security perceptions and behaviors: Polycontextual contrasts between the United States and China, MIS Q. 40 (1) (2016) 205–222.

[98] M. Karjalainen, S. Sarker, M. Siponen, Toward a theory of information systems security behaviors of organizational employees: a dialectical process perspective, Inf. Syst. Res. 30 (2) (2019) 687–704.

[99] J.G. Proudfoot, J.L. Jenkins, J.K. Burgoon, J.F. Nunamaker, More than meets the eye: how oculometric behaviors evolve over the course of automated deception detection interactions, J. Manag. Inf. Syst. 33 (2) (2016) 332–360.

[100] S.J. Pentland, N.W. Twyman, J.K. Burgoon, J.F. Nunamaker, C.B.R. Diller, A video-based screening system for automated risk assessment using nuanced facial features, J. Manag. Inf. Syst. 34 (4) (2017) 970–993.

[101] D. Gefen, I. Benbasat, P. Pavlou, A research agenda for trust in online environments, J. Manag. Inf. Syst. 24 (4) (2008) 275–286.

[102] Y. Ding, N. Luktarhan, K. Li, W. Slamu, A keyword-based combination approach for detecting phishing webpages, Comput. Secur. 84 (2019) 256–275.

[103] O.K. Sahingoz, E. Buber, O. Demir, B. Diri, Machine learning based phishing detection from urls, Expert Syst. Appl. 117 (2019) 345–357.

[104] E. Strickland, How Ibm Watson Overpromised and Underdelivered on Ai Health Care, 2019. Available at https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care. (Accessed 10 May 2020).

[105] D. Gefen, E.E. Rigdon, D. Straub, An update and extension to sem guidelines for administrative and social science research, MIS Q. 35 (2) (2011) Iii–Xiv.

Yan Chen is an associate professor at the Florida International University. She has received her doctoral degree from University of Wisconsin-Milwaukee. Her research focuses on information security, online fraud, security management, privacy, and e-business. She has published more than 30 referred research papers in academic journals and conference proceedings, including *MIS Quarterly*, *Journal of Management Information Systems*, *Journal of the Association for Information Systems*, *Information & Management*, and others. She is a recipient of research scholarships and best paper award nominees. She is a member of the Association for Information Systems and has been serving as a reviewer for many IS journals and conferences, including *MIS Quarterly*, *Information Systems Research, Journal of Management Information systems*, *Decision Sciences*, *Information & Management*, and others.

Fatemeh Mariam Zahedi is University of Wisconsin-Milwaukee Distinguished Professor Emerita at the Sheldon B. Lubar School of Business, University of Wisconsin-Milwaukee. She received her doctoral degree from Indiana University. Her research focus has been on IT at the service of individuals. Her present areas of research include web-based systems design and issues including trust, security, privacy, culture, loyalty, personalized intelligent interface, web-based healthcare, and web analytics for health. She has served as senior editor and associate editor of *MIS Quarterly*, editorial board of *JMIS*, and AE of *ISR*. She has published more than 120 referred papers in premier journals and conferences, including *MIS Quarterly, Information Systems Research, Journal of Management Information Systems, Management Science, DSS, Information & Management, IEEE Transactions on Software Engineering, Operations Research, IEEE Transactions on Systems, Man, and Cybernetics, IIE Transactions, and Review of Economics and Statistics,* and others. She has been the PI of grants funded by NSF and other agencies. She is the author of two books in *Quality Information Systems* and *Intelligent Systems for Business: Expert Systems with Neural Network*. She has received several research, teaching, and best paper awards. Her work has been featured on TV and in print media. The list of Professor Zahedi's publications is available on her Google Scholar profile.

Ahmed Abbasi is the Joe and Jane Giovanini Endowed Chaired Professor in the Department of IT, Analytics, and Operations in the Mendoza College of Business at the University of Notre Dame. He attained his B.S. and MBA degrees from Virginia Tech and a Ph.D. from the Artificial Intelligence Lab at the University of Arizona. His research interests relate to human-centered analytics, text mining, health, and security. He has published over 90 articles in top journals and conferences, including *MIS Quarterly*, *Journal of Management Information Systems*, *ACM Transactions on Information Systems*, *IEEE Transactions on Knowledge and Data Engineering*, and *IEEE Intelligent Systems*. His projects on cyber security, health analytics, and social media have been funded by the National Science Foundation and various industry partners, including AWS, Microsoft Research, and Oracle. He received the IBM Faculty Award, IEEE Technical Achievement Award, and INFORMS Design Science Award for his research on novel applications of machine learning. He has also received best paper awards from *MIS Quarterly*, the *Association for Information Systems*, and the *Workshop on Information Technologies and Systems*. He serves as senior editor at *Information Systems Research* and associate editor for *ACM Transactions on MIS* and *IEEE Intelligent Systems*. His work has been featured in several media outlets, including the Wall Street Journal, Associated Press, WIRED, and CBS.

David Dobolyi is an assistant research professor in the Mendoza College of Business at the University of Notre Dame. He received his Ph.D. in Cognitive Psychology from the University of Virginia, and his primary research interests involve predictive analytics, computer vision, and behavioral experiments, with recent applications including cybercrime and health. He has published in top journals including *Science* and the *Journal of Management Information Systems*, and his publications span a broad range of topics including reproducibility in science and the fusion of psychometric and secondary data for user modeling.