“I personally relate it to the traffic light”

a user study on security & privacy indicators in a secure email system committed to *privacy by default*

Borce Stojkovski
SnT, University of Luxembourg
borce.stojkovski@uni.lu

Gabriele Lenzini
SnT, University of Luxembourg
gabriele.lenzini@uni.lu

Vincent Koenig
COSA, University of Luxembourg
vincent.koenig@uni.lu

ABSTRACT

Improving the usability and adoption of secure (i.e. end-to-end encrypted) email systems has been a notorious challenge for over two decades. One of the open questions concerns the amount and format of information that should be communicated to users to inform them of the security and privacy properties with respect to different messages or correspondents. Contributing to the ongoing discussion on the usability and effectiveness of security and privacy indicators, particularly in the context of systems targeting non-expert users, this paper sheds light on users' evaluation of traffic light-inspired indicators, as a metaphor to represent different privacy states and guarantees, provided by a new system for email end-to-end encryption called p≡p. Using a mixed-methods approach, based on input gathered from 150 participants in three online studies, we highlight the pros and cons of the traffic light semantic in p≡p's context and beyond, and discuss the potential implications on the perceived security and use of such systems.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**; • **Human-centered computing → User studies**;

KEYWORDS

usable security, privacy engineering, privacy indicators, secure email, user studies

ACM Reference Format:
Borce Stojkovski, Gabriele Lenzini, and Vincent Koenig. 2021. “I personally relate it to the traffic light”: a user study on security & privacy indicators in a secure email system committed to *privacy by default*. In *The 36th ACM/SIGAPP Symposium on Applied Computing (SAC '21), March 22–26, 2021, Virtual Event, Republic of Korea.* ACM, New York, NY, USA, Article 4, 10 pages. https://doi.org/10.1145/3412841.3441998

1 INTRODUCTION

With the proliferation of information and communication technology, we have transferred many real-world concepts into the digital realm. While the degree of resemblance between the user interface representation and the real-world counterparts can vary among different systems, the basic idea is to enable users to interact with and via a system using concepts that they can recognize.

SAC '21, March 22–26, 2021, Virtual Event, Republic of Korea

ACM ISBN 978-1-4503-8104-8/21/03.
https://doi.org/10.1145/3412841.3441998

One such example is the notion of traffic lights. In the real world, they are universal signalling devices that control and regulate the flow of traffic, and whose mode of operation is something that we have learnt to interpret within a specific context (e.g. as drivers, pedestrians etc.). The digital counterparts, either derived directly or inspired by the traffic light semantic, can be found across various computer systems, and, in particular, those that are security and privacy critical. For instance, user interface representations of risks, alerts, and warnings often follow the *Red*/*Amber*/*Green* model.

While all those systems have the traffic light semantic in common, the represented concepts and their interpretation are often context-dependent and system-specific. The possibility to deploy the traffic light semantic within a specific context is fundamentally related to Ashby's ‘Law of Requisite Variety’ [4], which here essentially means the following: a traffic light system has a finite number of states that it can represent. Consequently, a system that seeks to deploy the traffic light semantic when communicating security or privacy critical information to users needs to take into account the number of signals it can effectively send to its users.

In this regard, we were motivated to investigate whether the traffic light semantic can be effectively utilized in systems for end-to-end encryption of email. We framed our investigation within a relatively new and promising secure email system that has *privacy by default* both as a guiding principle and as a key selling point. This system, called “Pretty Easy Privacy”, or p≡p[1], argues to have the traffic-light semantic at its core as a “clear and easily understandable presentation” [27] of the different privacy states that messages and communication peers can have.

On the whole, our work contributes to the ongoing discussion on the usability and effectiveness of security and privacy indicators, which in the secure email context has received relatively less attention. Through an empirical study involving 150 participants, we shed light on users' evaluation of traffic light-inspired indicators, used as a metaphor to represent the privacy states and guarantees that a secure email system can provide.

More concretely, our findings provide direct feedback to the developers of p≡p by *(i)* reporting on user research that probes p≡p's reworked approach to utilizing traffic-light inspired indicators, and *(ii)* studying the adequacy of such indicators to represent desired privacy states, from a users' perspective. We also *(iii)* inform how user evaluation of an indicator differs when examined alone or as part of an indicator set, highlighting the complex relationship between different indicator cues and their interpretation with respect to security and privacy. This is relevant beyond our use case.

The next sections present the frame of our investigation, the methodology, results and analysis. We conclude with a discussion and future research directions.

[1] https://www.pep.security/, accessed on October 5, 2020

2 CONTEXT AND RELATED WORK

In this paper, we focus on systems for *secure email* i.e. those that offer end-to-end encryption, which is the highest level of protection possible. Despite the lack of widespread adoption (as discussed in 2.3), we believe that research in this topic is not only of particular significance for individuals, marginalized communities or disadvantaged groups that rely on these technologies to protect their privacy, but also for professionals and businesses that strive to meet legal requirements with respect to data protection, such as the GDPR.

In addition to the issues of unsecured links and message forgeries, end-to-end encryption also addresses the problem of untrusted servers [35]. Secure email, thus, provides the guarantees of *confidentiality*, *integrity*, *authenticity*, and *non-repudiation* [24]. From an end-user's perspective, this means: i) there are mechanisms in place that protect the email content from being read by entities other than the intended recipients; ii) the contents of the message are received just as they were sent; iii) a recipient can verify whether a message was sent by a party who is in possession of a specific (private) key; iv) a recipient of a message can convince others that the message was sent by a party in possession of a specific (private) key i.e. the sender cannot successfully deny that she sent the message.

While these guarantees are achieved through well-established crypto primitives, such as public-key cryptography and digital signatures, communicating in a secure and private fashion is a broader socio-technical challenge. For instance, digital signatures provide authenticity in the true sense, as long as the recipient is certain that the sender is really who they say they are (i.e. the recipient holds and trusts a public key that corresponds to the private key which signed the message).

Thus, it quickly becomes evident that secure email systems need to discern many different *states* that correspond to security and privacy concepts being met or violated (e.g. is a message encrypted, is it signed, is the key trusted, mistrusted or unknown, etc.). How granular should this distinction be from an end-user's perspective and how best to convey this to users?

Before we discuss related work on secure email and security indicators in that context more deeply, we provide an overview of the theoretical concepts underpinning security indicators, warnings and risk communication in computer security.

2.1 Human factors and warning research

A security indicator can be understood as a medium through which security experts communicate the results of an analysis of a security-sensitive system to its users. Its role is, thus, to inform the users of the results of this analysis with the implicit expectation that it helps users understand the implications of the result, as well as the actions that they should or should not take. In other words, its purpose is also to influence the behavior of the user. Finally, it also serves as a reminder to help a user — who may be knowledgeable of a potential threat, the associated consequences, or would otherwise exhibit the appropriate behavior — become aware of the security status of the system at a crucial point in time.

Given the numerous risks and hazards that are inherent in computer security and privacy systems, security indicators bear many of the same characteristics as warnings, which are the third line of defense in the *hazard control hierarchy* [23]. The sequence of stages through which warning information flows can be described using the C-HIP model [46]. In a nutshell, for a warning to be successful, it must not only capture attention and be understood, but also align with existing beliefs and attitudes and motivate users to comply [47]. Furthermore, there are a number of design and non-design factors that influence warning effectiveness, in particular regarding attention (i.e. noticing and encoding a warning) and compliance (i.e. costs-benefit trade-off decisions) [23].

In contrast to early views which considered emotion and reason at odds with each other, research shows that cognition and emotion are closely intertwined and to a large extent cooperative [26]. Emotions, thus, have impact on attention, working memory, information processing and decision-making. Besides, there are individual as well as cultural differences in how people perceive, process, and behave toward affective stimuli [26].

There are significant methodological challenges associated with the evaluation of warning effectiveness. Nevertheless, testing by means of exposing the warning to a representative sample of the target audience and assessing specific properties (e.g. noticeability, readability, comprehension, behavioral intention and behavioral compliance), can be an effective approach that should be integrated into the warning design process [47].

2.2 Risk communication in computer security

In the computer security setting, *explanations* are thought to bridge the gap between the *actual* and *perceived* security [29]. There exists, therefore, a clear need to provide appropriate feedback about security and communicate risks, so that users can make informed decisions [8, 44]. To this end, visual feedback mechanisms have been proposed to help users operate security or defend themselves from the growing number and sophistication of attacks online.

However, research shows that computer warnings and security indicators have oftentimes been ineffective [10, 13, 22]. Users either ignore [16, 41] or do not even take notice of security indicators [38, 48], they do not understand them [14], or they underestimate the associated risks or are completely unaware of the risks [9, 14]. Users ignore warnings as they become desensitized by frequent exposure and false alarms [22]. Interruptions substantially impact alert disregard [20], while habituation (the diminishing of attention because of frequent exposure) seems to be largely obligatory as a result of how the brain processes familiar visual stimuli [3].

Mental models have been proposed as a method to improve communication to users about computer security risks [11], as well as an approach to getting insights into how users perceive and respond to computer alerts. Bravo-Lillo and colleagues highlighted that advanced and novice users observed different sets of cues, had a different interpretation of the underlying risks, and exhibited different responses [10]. As risk communication is mainly designed by computer scientists, it is often influenced by mental models of experts [9] which is problematic for many systems that rely on a "human in the loop" to perform security-critical functions [12].

Over the years, significant improvements to both warning adherence [2] and comprehension [17] have been reported in the context of web browsers, wherein much research was conducted. In contrast, the question of security indicators within systems for secure email has received relatively less attention.

2.3 Usability and adoption of secure email

Unlike secure instant messaging applications, end-to-end email encryption has failed to achieve widespread adoption even though solutions based on two of the most popular standards, PGP and S/MIME, have been around for more than two decades.

The seminal work by Whitten and Tygar [45] showed significant usability problems with the existing PGP client at the time. While a subsequent study using an updated version of PGP showed similar results [39], combining the idea of Key Continuity Management (KCM) with S/MIME [18] suggested that automatic key generation and management was more usable than the manual key management in the original study.

Research suggests a strong user preference for encryption tools that offer a tight and seamless integration with users' existing email systems over standalone encryption software [5, 32, 34]. Studies on the implications of automatic vs manual encryption on usability and trust is mixed. While some argue that trust is reduced when secure systems hide from users how they provide security [32, 34], other results indicate that user trust was not impacted by the transparency of encryption tools [5]. Despite having an equal motivation for protecting their communications, two different user groups can have sufficiently different requirements, that they may require entirely different tools for email encryption [25]. Furthermore, encrypted email may be unhelpful or worse, if it provides a false sense of security to certain groups that have special requirements [25]. Designers should, thus, explain the security properties that encryption tools offer [7], whereby inline, context-sensitive tutorials and streamlined onboarding appear to be essential [33].

In addition to poor interface design choices, key management difficulties and mistaken mental models, researchers have identified social and cultural norms as factors that contribute towards non-adoption of email encryption too [18, 19, 30]. Similarly, usability might not necessarily be the primary obstacle to the adoption of secure communication tools, but rather fragmented user bases, lack of interoperability and low quality of service [1].

2.4 Indicators in secure email

Indicators in the context of secure email are very much linked to metaphors, a number of which have been proposed in an effort to help users understand the underlying complexity of PKI [6, 31, 43]. Lausch, Wiese and Roth reviewed existing indicators used in secure email systems, and performed a comparative study to identify the ones best suited to represent the concept of email security [24]. The findings highlighted that postcards, mail envelopes, and a torn envelope emerged as promising candidates on par with the dominant padlock for signalling the encryption and integrity states.

Garfinkel and Miller investigated the effects of indicators in relation to security threats, such as social engineering and new-identity attacks [18]. They pointed out that users would need to occasionally face trust decisions, and they defined color-codes for security indicators depending on different situations. A *yellow* indicator would appear if a digitally signed message is received from a particular address for the first time; a *green* one for subsequent messages with the same key; a *red* one if a different key is used for that address (with the possibility for users to override the code); and a *gray* one if the message is unsigned.

A similar traffic-light inspired approach can be found in p≡p, as described next. A preliminary investigation of p≡p's indicator effectiveness hinted at potential problems, which might have helped trigger a discussion at p≡p on their design choices [40].

3 P≡P

As a use-case for our investigation, we employed p≡p whose underlying crypto relies on OpenPGP libraries, and whose cryptographic protocols were investigated and found to be secure [36].

Based on opportunistic security [15], p≡p positions itself as technology for secure and private communication that has usability as a key motivation or goal [28]. Targeting primarily non-expert users, p≡p's approach is not to confront its users with technical jargon around cryptography. It automates the majority of user-related operations, e.g. key management, key discovery, private key handling etc. It has been designed with functionality, security and privacy considerations, such as interoperability, minimal configuration, and in particular, no trusted servers. It can be used for communication in both encrypted and plain text formats, with people that do or do not use p≡p or other encryption software. The desktop distributions of p≡p integrate into Outlook and Thunderbird, whereas the iOS and Android distributions work as standalone clients.

p≡p Privacy States. As per p≡p's documentation [27, 28], and as summarized in Table 1, the system differentiates between 13 internal privacy rating states, which are assigned corresponding number codes, color codes and labels. Captions and explanations are provided for a subset of the states that are visible in the user interface (UI). While p≡p assigns a privacy status to messages or correspondents automatically based on several factors, certain states i.e. *Mistrusted* and *Secure & Trusted*, can be reached only in combination with user interaction (i.e. users have to explicitly confirm the correspondent's authenticity in the p≡p client). The implicit expectation is that users will seek to communicate in the *Secure & Trusted* state as it guarantees the highest protection possible.

p≡p Security and Privacy Indicators. The mapping of the internal privacy states to the corresponding UI elements results in a set of indicators that follow the traffic light semantic. p≡p accounts for color-blindness in potential users by additionally providing a distinctive shape with each indicator.

The default visual indicators, as advertised on the p≡p website or implemented in p≡p for Outlook (ver. 1.1), can be seen in Fig. 1a. While p≡p promotes only three color codes i.e. a *red*, *yellow*, and *green* indicator, the one with color code 0 (no color) can effectively be seen as a fourth indicator in *gray* when implemented in the UI.

The only related study [40] that we are aware of, examined p≡p's default design choices by asking prospective users which of those 4 visual icons they would associate with each of the different UI labels and explanations[2]. There, we discovered that the icon displayed by p≡p matched the association made by the test participants for only 4/10 states in the case of UI labels, and 3/10 states in the case of UI explanations. Furthermore, for the privacy state *reliable*, none of the participants (0%) matched the yellow triangle to the UI label *Secure*, which is the icon and label combination in p≡p.

[2]UI explanations are not provided in the documentation, but can be extracted from the source files of the p≡p distributions.

Rating Code	Rating Label	Color Code	Color Label	UI Label
-3	under attack	-1	red	Under Attack
-2	broken	-1	red	Broken
-1	mistrust	-1	red	Mistrusted
0	undefined	0	no color	Unknown
1	cannot decrypt	0	no color	Cannot Decrypt
2	have no key	0	no color	-/-
3	unencrypted	0	no color	Unsecure
4	unencrypted for some	0	no color	Unsecure for Some
5	unreliable	0	no color	Unreliable Security
6	reliable	1	yellow	Secure
7	trusted	2	green	Secure & Trusted
8	trusted and anonymized	2	green	-/-
9	fully anonymous	2	green	-/-

Table 1: Overview of p≡p's internal privacy rating codes, color codes, color labels and UI labels

(a) Old version **(b) New version**

Figure 1: Security and Privacy indicators in p≡p

p≡p Indicators – Revisited. Interacting with updated versions of p≡p and contacting the developers, we learned that p≡p has updated the indicator shapes, while keeping the color codes and traffic light metaphor. In the new version, shown in Figure 1b, *Mistrusted* is represented with a red triangle, *Secure* with a yellow/amber circle, and *Secure & Trusted* with a green shield pointing downwards. As per the Android onboarding tutorial (ver. 1.1.008), there is no *gray* indicator, and it appears to be left out in the UI.

4 OUR STUDY

Coupling our research motivation to a representative use-case, i.e. a real-world system that aspires to achieve *privacy by default*, we sought to conduct a basic, yet fundamental investigation on the use of traffic light indicators for communicating security and privacy information to users in a secure email context. Aspiring to highlight the importance of early user research in the development process of privacy-enhancing tools, we were driven by the following:

- How do we compare different design alternatives that try to convey specific information via the traffic light semantic?
- Which of the proposed indicators do end-users find appropriate for the designated privacy states?
- Does the perception about a traffic light indicator change when it is considered as part of an indicator set, rather than individually?

To this end, we formulated the following hypotheses:

H1 — H3: The majority of participants would select the new versions of the icons over the old versions for each of the three privacy states *Mistrusted*, *Secure*, and *Secure & Trusted*.

H4: The majority of participants would express agreement that the new version of the icon (red triangle) is a good representation of the text *Mistrusted*.

H5: The majority of participants would *not* express agreement that the new version of the icon (yellow circle) is a good representation of the text *Secure*.

H6: The majority of participants would express agreement that the new version of the icon (green shield pointing downwards) is a good representation of the text *Secure & Trusted*.

H7: Onboarding has a positive effect i.e. participants exposed to a priming screen displaying the whole indicator set, express higher agreement scores versus non-primed participants across all three states.

5 METHODOLOGY

In order to test our hypotheses we conducted three independent, within-subject experiments, Study A, B and C, as described below.

Recruitment. The study participants were recruited via the platform Prolific[3]. Given that the icons in the investigation had different colors, we restricted the participation to those that could see color normally. In total, 152 participants were recruited, thereof 150 were eligible and taken into consideration (50 per study). To ensure independence of the experiments and exclude any accidental participant overlap, the studies were conducted sequentially and all participants were "blocked" for further recruitment.

Survey. The experiments were conducted online. We administered one survey per study via Qualtrics.

Ethics. Our study was approved by our organization's ethics review panel, and we obtained informed consent from all subjects.

Compensation. The participants were informed that it would take them about 3 minutes to complete the survey. They were compensated £0.25 for their participation, which corresponds to Prolific's fair rewarding practice of at least £5.00 ($6.50) per hour.

5.1 Experiment protocol

Full versions of the study surveys can be found in the Appendix.

Information and Consent. At the beginning of all studies, the participants were prompted that the survey is part of an investigation that aims to research and improve the user experience of products and systems for secure messaging, in particular secure email. We informed them that we are interested in understanding how icons can be used for communicating different levels of privacy for messages exchanged in such systems. After consenting to take part in the study and confirming that they see color normally, depending on which study the participants were part of, they were shown three consecutive questions, as described below.

Study A (H1 — H3). First we wanted to investigate how participants would evaluate or score the two sets of icons with respect to the privacy states that they are supposed to represent. The purpose

[3]https://www.prolific.co/, accessed on October 10, 2020

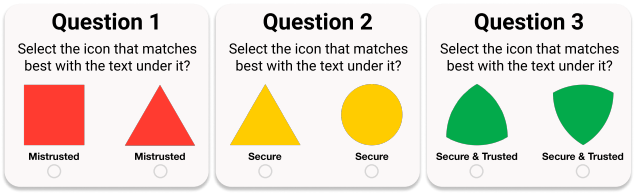


Figure 2: The three preference questions shown in Study A

was to understand among prospective p≡p users, the preference between the old and new versions of the icons designated to represent three different privacy states i.e. *Mistrusted*, *Secure*, and *Secure & Trusted*. Preference, here, refers to the selection of one icon alternative over the other, based on the perceived fitness of the icon with the corresponding privacy state, labeled under each icon.

Fig. 2 features the question set shown to each participant, each shown on a separate page. To counterbalance possible biases, the order of the questions and answer options was randomized.

Study B (H4 — H6). Next we wanted to find out how strong is the presumed fitness between the proposed icon and the corresponding label. In other words, while a new version of an icon might be better than its predecessor, it does not mean that it is a good representation for the privacy concept that it is supposed to convey. Each Study B participant was shown, in a randomized fashion, a set of three questions, asking her to state on a 7-point rating scale how much she agrees or disagrees with the statement that the displayed icon is a good representation of the text under it. Only the new versions for each of the three privacy states were displayed.

Study C (H4 — H7). Finally, we conducted a follow-up investigation almost identical to Study B, whereby the traffic light semantic was made more explicit. This was done to see if there is an effect of the onboarding on the evaluation of the individual fitness of the indicators. Thus, the crucial difference was the inclusion of an onboarding screen, where all three icons were displayed all-together on a page, before the Likert item questions were randomly shown to the participants, as in Study B. Another minor change was that along with each rating question, a non-mandatory free-entry question *"Why do you think so?"* was also shown. This was done in order to gather additional input and try to understand, whenever possible, what reasoning backed the fitness scores that the participants gave.

Demographics. At the end of all experiments, there was a demographics section where we inquired if our participants had a computer science / technical background and whether they had ever used systems for end-to-end encryption (E2EE) of email. Those that affirmed, were further asked to name the systems that they use or had used in the past. We asked this to establish any skewness of our sample towards privacy-aware and tech-savvy users.

6 RESULTS AND ANALYSIS

6.1 Participants

Table 2 provides an overview of the main participant demographics. Our sample consists of participants that have different technical skills and experience with systems for secure email. PGP and Protonmail were mentioned most frequently as tools/systems that they use or have used in the past. No participant mentioned p≡p.

Demographics	Study A	Study B	Study C
Female	28 (56%)	22 (44%)	18 (36%)
Male	19 (38%)	28 (56%)	32 (64%)
No attribution	3 (6%)	0	0
Average age	28	32	29
Age range	[18 – 46]	[18 – 63]	[18 – 69]
English as first language	28 (56%)	21 (42%)	13 (26%)
Student status	19 (38%)	19 (38%)	13 (26%)
CS / tech background	16 (32%)	22 (44%)	14 (28%)
Use of E2EE systems	13 (26%)	14 (28%)	10 (20%)

Table 2: Participant demographics. N=50 for each study.

6.2 Quantitative analysis

6.2.1 Study A. The proportions of icon preference were estimated using exact binomial tests. The results, displayed in Fig. 3 and Table 3, confirm H1 — H3 that, for each state, the majority of participants would select the new versions of the icons over the old ones.

The new version was selected as the one that better matches with the text under it in 40/50 times in the case of the *Mistrusted* and *Secure* privacy states. For *Secure & Trusted*, it was 47/50 times. The confidence intervals for the new versions are way above chance performance of Π_0 =.5 (p <.001), confirming H1, H2 and H3.

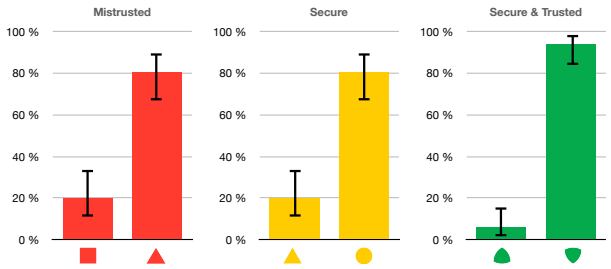


Figure 3: Study A - Proportions of frequencies of the two icon versions per privacy state

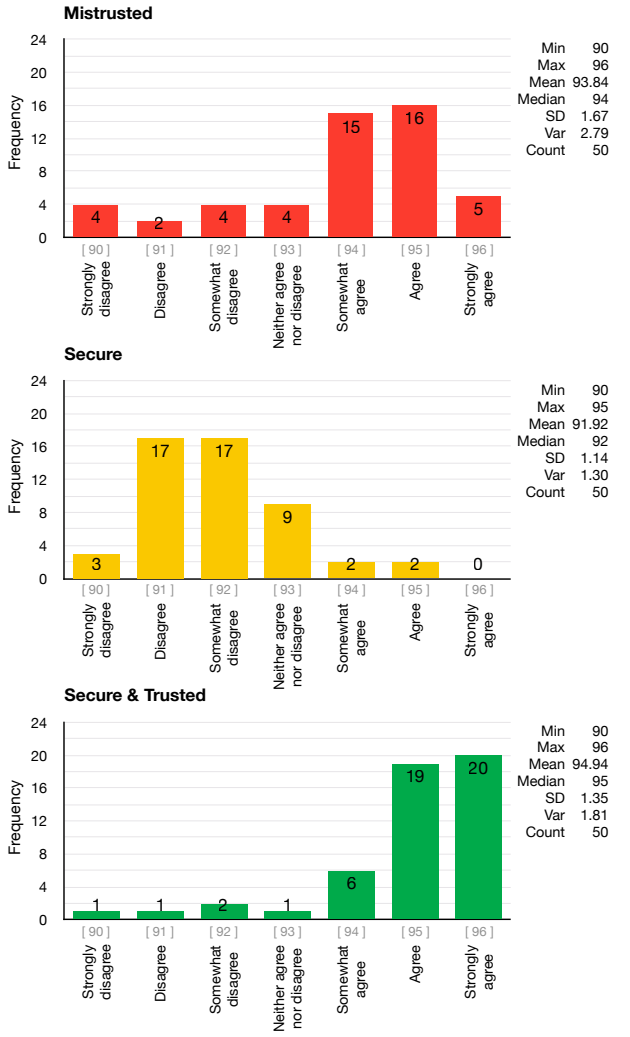
Icon	Version	#	Count	Percent	Proportion	95% CI*	Mean	SD	Var
■	Old	1	10	20 %	$P_{M\text{-}1}$ 0.2	[.1124, .3304]	1.80	.404	.163
▲	New	2	40	80 %	$P_{M\text{-}2}$ 0.8	[.6696, .8876]			
	Total		50	100 %	1				
▲	Old	1	10	20 %	$P_{S\text{-}1}$ 0.2	[.1124, .3304]	1.80	.404	.163
●	New	2	40	80 %	$P_{S\text{-}2}$ 0.8	[.6696, .8876]			
	Total		50	100 %	1				
	Old	1	3	6 %	$P_{ST\text{-}1}$ 0.06	[.0206, .1622]	1.94	.240	.058
	New	2	47	94 %	$P_{ST\text{-}2}$ 0.94	[.8378, .9794]			
	Total		50	100 %	1				

*CI method: Wilson Score interval

Table 3: Study A - Statistics and Frequency Table

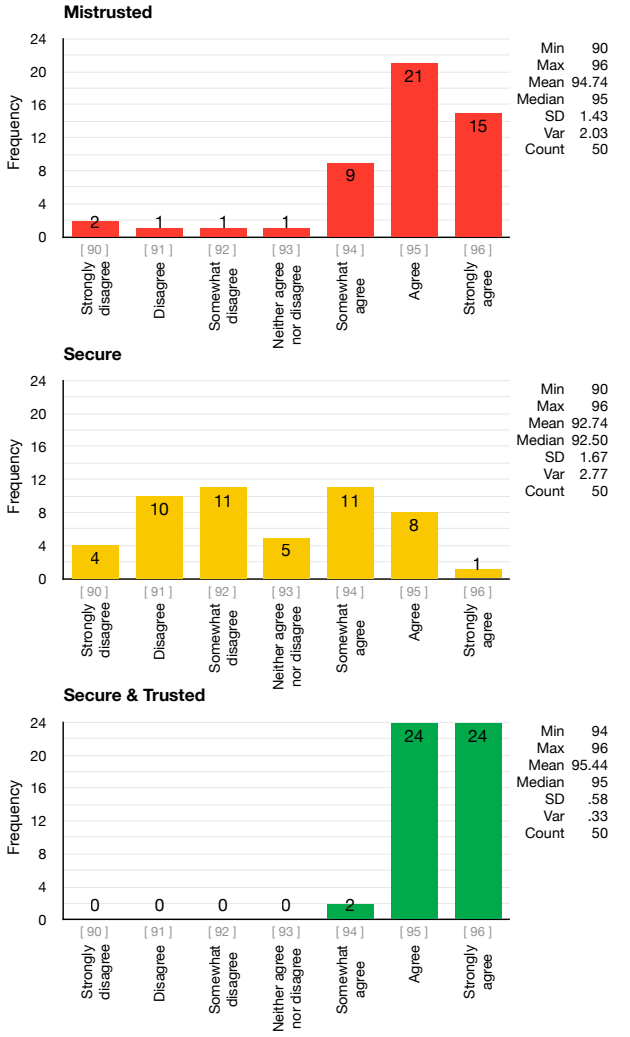
6.2.2 Study B and Study C. Figures 4a and 4b show the distributions of the responses to the three rating questions from Study B and C. As visible in the figures and summarized in Table 4, the majority of participants express agreements that the icon is a good match for the text for the *Mistrusted* and *Secure & Trusted* privacy states in both studies. These high agreement scores are in contrast to the ones expressed for the privacy state *Secure*.

State	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree	Total
▲	4	2	4	4	15	16	5	50
M	8 %	4 %	8 %	8 %	30 %	32 %	10 %	100 %
●	3	17	17	9	2	2	0	50
S	6 %	34 %	34 %	18 %	4 %	4 %	0 %	100 %
♥	1	1	2	1	6	19	20	50
S&T	2 %	2 %	4 %	2 %	12 %	38 %	40 %	100 %



(a) Study B (i.e. without onboarding)

State	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree	Total
▲	2	1	1	1	9	21	15	50
M	4 %	2 %	2 %	2 %	18 %	42 %	30 %	100 %
●	4	10	11	5	11	8	1	50
S	8 %	20 %	22 %	10 %	22 %	16 %	2 %	100 %
♥	0	0	0	0	2	24	24	50
S&T	0 %	0 %	0 %	0 %	4 %	48 %	48 %	100 %



(b) Study C (i.e. with onboarding)

Figure 4: Frequency histograms for participant responses to the three main rating questions in Study B (left) and Study C (right). The 7 response categories are ordered and assigned numerical codes [90 to 96] on the x-axis.

Mistrusted: The combined (Study B and Study C) agreement score denotes that the percentage of participants that either *Strongly agree*, *Agree* or *Somewhat agree* that the red triangle is a good representation for *Mistrusted* is **81% [72.22, 87.49]**. This is way above the benchmark of 50% of participants, thus confirming Hypothesis 4. Onboarding appears to be associated with an increase of the aggregated agreement score of 18 percentage points, from 72% in Study B (no onboarding screen) to 90% in Study C (with onboarding).

Secure: The percentage of participants in Study B that express agreement that the yellow circle is a good representation for the text *Secure* is 8% [3.15, 18.84]. Providing an onboarding screen is

Study	State	Agreement	95% CI
B	Mistrusted	72%	[58.33, 82.53]
C	Mistrusted	90%	[76,84, 95.65]
Combined	Mistrusted	81%	[72.22, 87.49]
B	Secure	8%	[3.15, 18.84]
C	Secure	40%	[27.61, 53.82]
Combined	Secure	24%	[16.69, 33.23]
B	Secure & Trusted	90%	[78.64, 95.65]
C	Secure & Trusted	100%	[92.87, 100]
Combined	Secure & Trusted	95%	[88.82, 97.85]

Table 4: Aggregated percentage of participants that *Somewhat agree, Agree* or *Strongly agree.* (*CI method: Wilson score*)

associated with an increase of the aggregated agreement score of 32 percentage points to 40%. Nevertheless, the 95% confidence interval around the percentage of primed participants who would express agreement is between 27.61% and 53.82%, denoting that we do not have convincing evidence that the majority of participants would be in agreement. The combined (Study B and Study C) agreement score for the *Secure* privacy state is **24% [16.69, 33.23]**, way below the benchmark of 50%, thus confirming Hypothesis 5.

Secure & Trusted: The highest agreement scores were expressed for the green shield icon and the *Secure & Trusted* label. The combined (Study B and Study C) agreement score is **95% [88.82, 97.85]**, thus confirming Hypothesis 6. Onboarding is associated with an increase of 10 percentage points, from an already high 90% to 100%.

6.2.3 Differences in agreement scores across different categories of onboarding (between-subject). Mann-Whitney U tests were run to determine if there were differences in the agreement score between study participants that were exposed to an onboarding screen (primed) and not i.e. Study C *versus* Study B.

Mistrusted: Distributions of the agreement scores for both groups were similar, as assessed by visual inspection. Median agreement score was statistically significantly higher for primed participants (95) than for non-primed ones (94), $U = 1712$, $z = 3.308$, $p = .001$.

Secure: Distributions of the agreement scores for primed and non-primed participants were not similar, as assessed by visual inspection. Agreement scores for non-primed participants (mean rank = 43.48) were statistically significantly lower than for primed participants (mean rank = 57.52), $U = 1601$, $z = 2.480$, $p = .013$.

Secure & Trusted: Distributions of the agreement scores were similar, as assessed by visual inspection. Median agreement score was not statistically significantly different between primed and non-primed participants, $U = 1468$, $z = 1.645$, $p = .100$.

While participants in Study C expressed higher agreement scores than those in Study B, the Mann-Whitney U tests above indicate that this difference was statistically significant only for the *Mistrusted* and *Secure* privacy states, thus Hypothesis 7 is only partially confirmed. It is important to note, however, that the agreement score in Study B was already high at 90%.

6.2.4 Differences in agreement scores based on technical background (between-subject). Further Mann-Whitney U tests were run to determine differences among participants with and without a technical / computer science background. The agreement scores were not statistically significantly different.

- *Mistrusted*: $U = 1249.5$, $z = .727$, $p = .467$.
- *Secure*: $U = 1252.5$, $z = .740$, $p = .460$.
- *Secure & Trusted*: $U = 1179$, $z = .212$, $p = .832$.

6.3 Qualitative analysis

Themes	Theme frequency per state			
	M	S	S&T	**Total**
Indicator characterization				
- The color is adequate	24	4	24	**52**
- The color is not adequate	2	18	0	**20**
- The shape is adequate	8	2	20	**30**
- The shape is not adequate	7	9	2	**18**
- The indicator is confusing	0	9	0	**9**
Indicator interpretation				
- Traffic light semantic	2	5	7	**14**
Evocation of feelings				
- Sense of security	0	3	22	**25**
- Sense of reassurance	1	1	14	**16**
- Sense of caution	10	12	0	**22**
- Sense of danger	15	1	0	**16**

Table 5: Overview of the most frequent themes emerging from the data during the qualitative analysis.

Input to the optional *"Why do you think so?"* question in Study C, was provided by 33 participants for the *Mistrusted*, 37 participants for the *Secure*, and 32 participants for the *Secure & Trusted* state.

Based on this data, the first author performed inductive category formation in consultation with the other authors. Table 5 provides an overview of the main themes identified per privacy state.

6.3.1 Mistrusted. As summarized in Table 5, column (M), and as visible from the following verbatims, participants tend to agree that the color of the *Mistrusted* indicator is appropriate, but they are divided when it comes to the shape of the indicator.

- *"the triangle does not make me think mistrusted or problematic. Red is a good choice tho."* (P107)
- *"It is bold and makes you stop and think. Red is a good representation of danger."* (P111)
- *"the red colour is a good warning sign, the colour is powerful so would catch your attention."* (P112)
- *"Colour is adequate, geometrical form could be better"* (P115)

As hinted in the verbatims above, the indicator was mostly associated with *danger* and *caution*.

6.3.2 Secure. The most frequent themes under the (S) column in Table 5 and the representative verbatims below provide first insights as to why the indicator received a poor overall score in the 7-point rating question. In most cases, the color choice for the indicator was criticized for not being representative of the concept of *Security*:

- *"secure is usually in a green symbol."* (P102)
- *"the yellow represents a colour which is not secure nor unsecure, in my opinion."* (P105)
- *"color doesn't seem to scream safe to me."* (P108)
- *"the yellow color doesn't seems to be so secure at all."* (P144)

These were accompanied by comments of *doubt* and *confusion*:

- *"not so sure that this indicates security."* (P114)
- *"to some, the color could be misleading."* (P139)

Participants also voiced the inadequacy of the circle:

- *"The shape is not good. I'd like prefer yellow shield."* (P119)
- *"'Secure' usually indicates a shield icon should be used."* (P124)
- *"circle isn't a particularly distinctive symbol."* (127)
- *"it should be shield image, it looks safe, not circle."* (P140)

In fact, the large number of low agreement scores provided in the rating question can also be explained by the participant association of the indicator with a *Sense of caution* rather than a *Sense of security*:

- *"I associate it with the yellow light in traffic, that in my country means proceed with caution."* (P117)
- *"Yellow signals warning for me so I would not feel it is secure."* (P128)
- *"Shape isn't anything special, additionally yellow colour associate, as if something dangerous."* (P143)

In contrast, there was also a small number of participants who associated the indicator with *caution*, yet expressed positive agreement scores for the corresponding rating question:

- *"As yellow is like amber use with caution."* (P106)
- *"Yellow generally means caution."* (P116)
- *"I personally relate it to the traffic light, it is not dangerous but it does not tell me that I am sure."* (P120)

While we hypothesize that p≡p envisions users to interpret the *Secure* indicator as in the above three verbatims, our results suggest that this feeling of *"self-reflective security"* (which we discuss in Section 7) is evoked only in a minor proportion of the users.

6.3.3 Secure & Trusted. Given that the agreement scores for this state were only positive (see Figure 4b), most comments, were confirmations of the adequacy of the color and shape of the indicator. Participants mentioned positive associations, such as *Sense of security* (22 times), and *reassurance* (14 times). A reference to a *Traffic Light Semantic* was observed 7 times. Representative verbatims include:

- *"I feel like green is colour of safety and that shape looks kind of shield. All of it makes me feel really secure and trusted."* (P143)
- *"The shield shape and the green colour are a trustworthy and appear regularly on computer programs."* (127)
- *"The color and shape make me feel at ease. I am used to green meaning go from driving so perhaps that has something to do with it as well."* (P111)

Nevertheless, the challenge of representing *trust* and the insufficiency of the shield icon to represent both the concepts of *security* and of *trust* was highlighted too:

- *"With a shield look to it, it looks like things should be okay to proceed. but I think it needs something else for the 'Trusted' part like a little start on it or a banner badge."* (P124)
- *"good representation for secure, but I think a different icon should be used for Secure & Trusted."* (P196)

6.4 Summary of key results

The new indicators are better. In comparison to the old version, participants find that p≡p's new visual indicators better correspond to the names for all three privacy states.

Better does not always mean good enough. Irrespective of their tech background, participants do not find p≡p's new indicator to be a good representation for the state *Secure*.

Onboarding has a positive effect. Exposing users to a priming screen with the whole indicator set impacts how users evaluate the individual indicators.

Onboarding is not a silver bullet. While users exposed to onboarding did find the indicator for the state *Secure* to be more fitting than those that were not exposed, the majority of them, nevertheless, disagreed that it is a good representation.

"Something is Rotten in the State of" *Secure*. Participant feedback clearly points to the color and shape of the indicator as not being adequate for the *Secure* state. Furthermore, the indicator evokes feelings of caution, rather than security.

Indicator shapes should not be downplayed. While overall the red indicator was evaluated as fitting, participant feedback hints at potential issues with the designated shape to represent the *Mistrusted* state. In the case of *Secure & Trusted*, it is not clear whether the green shield reflects both the concepts of *Security* and of *Trust*.

7 DISCUSSION

Coming up with effective indicators in systems for secure email is closely tied to these two user-related challenges: understanding and controlling secure email. The first deals with users' ability to recognize the security status for a particular message or correspondent that a system tries to communicate through a concept familiar to the user. The second deals with the amount of control that the user exerts over the system or is expected to contribute for the interaction to take place with the desired security outcome.

In view of the afore-mentioned complexities intrinsic to secure email, two options are available to systems that attempt to deploy traffic lights as means to communicate security information to their users: either reduce the variety in the environment (i.e. choose to communicate only a subset of the possible states); or increase the variety in the system (i.e. resort to additional "mechanisms" to communicate the desired states). While the number and relative ordering between the three states in p≡p allows for a direct mapping onto the indicators found in traffic lights, this is not as straightforward from a user's perspective, as our study shows.

What is the role of the yellow indicator? The key question boils down to: "What does p≡p want to communicate with the *Secure* privacy state"? A sense of security, a sense of caution, or both in order to accommodate for a range of threat models at the same time? We term this *self-reflective security*. As our investigation highlights, combining both is a daunting task. For experts and security-savvy users, it is immediately clear that in the *reliable* security state, users could be susceptible to a man-in-the-middle attack. Thus, p≡p attempts to signal this potential problem by suggesting a cautious approach using the yellow indicator. The name of the state in the UI, however, for the majority of users instills a sense of security, in contradiction to the visual indicator.

Our data suggests two avenues that could be explored to resolve the current discrepancy:

- Remove the secure association from the indicator, and communicate cautiousness more. This is a design choice i.e. if the *reliable* state is not secure, it should not be called *Secure*.
- Alternatively, if it is "secure enough" for non-expert users that do not have extensive threat models and already use other systems that offer the same or lower levels of protection (e.g. centralized E2EE instant messaging), then change the indicator to represent the concept of security (rather than caution). The system could still provide a hint on the indicator's position relative to other indicators in the set in order to denote that there might be a higher protection level, yet without unnecessarily sending mixed signals.

Implications on the perceived security. While discussing the preliminary analysis of this investigation with the p≡p developers, we realized that the reasoning behind the secure, yet cautious indicator can be traced back to their founding mission. p≡p is rooted in privacy activism, thus one of their core ideas is to "nudge" users to be more privacy-conscious. As research suggests, however, "greater familiarity, assuming no negative experiences in the past, may result in lower levels of perceived hazard and, in turn, less motivation to seek warning information" [23]. Meaning, users can be easily habituated if all goes well in the *Secure* state, and as verifying key fingerprints (or trustwords in p≡p's case) is neither a primary task, nor done frequently [42], users would be less likely to move to the next state with higher security guarantees i.e. *Secure & Trusted*.

This is problematic, because, on the one hand, users might have a false sense of security while still being susceptible to MITM attacks. On the other, for non-expert users without special security or privacy requirements, the perceived hazard is probably low, making the interaction with a system that sends mixed signals confusing, potentially impacting the usability and adoption.

Beyond the colors of traffic light indicators. Inclusive design aims to meet the needs of non-disabled and disabled users alike [21], which is of concern in the context of indicator and warning design too [47]. The introduction of shapes to improve the accessibility can come with side-effects, however, such as communicating additional information that may be in contradiction to the other cues.

Apart from colors and shapes, constitutive components of indicators are also the associated text labels. While substituting PKI jargon with non-technical terms is the right way forward for systems targeting non-expert users, one needs to bear in mind that such labels might carry connotations subjective to each user. In p≡p's case, do novice users understand what *Mistrusted* and *Trusted* refer to? Could it be the public key of the correspondent, the actual person or the contents of the message that they sent? In other words, misinformation, malicious links or malware attachments might also come via advanced E2EE systems, intentionally or unintentionally, even from people that we trust.

Exposing users to an indicator set, e.g. via "onboarding tutorials", can support them in positioning individual indicators with respect to the others in the set. This can potentially help users understand the ordinal location and by extension, any associated risks or security connotations, the system is trying to communicate to users, as long as they have a "correct" understanding of the extremes of the indicator spectrum. As such, traffic-light inspired indicators could be fit for purpose provided the intermediary state is neutral i.e. logically equidistant to the indicator spectrum ends.

Limitations. Our work is by no means exhaustive and comes with certain limitations that should be considered when interpreting the results and analysis of our study.

We conducted an investigation with hypothetical, prospective users out of context. While this removes prior bias that actual p≡p users might have had, it omits the context of use and situated interaction. Our approach can, nevertheless, be a useful first test of indicator recognition and information-scent.

We cannot exclude the possibility that the score deviations in Study B and C can be confounded by the possible extra differences of the participants, which is inherent to between-subject studies.

To investigate user opinion for each state, we used only one question. Nevertheless, given the importance of the number of steps for single-item measurements, we used a 7-step question as suggested by literature [37].

The study was grounded in one particular system. We argue, however, that the methods and insights are easily transferable to other privacy-enhancing systems that aim to or already employ the traffic light metaphor as a visual feedback mechanism.

8 CONCLUSION AND FUTURE WORK

User interfaces inspired by the traffic light semantic are omnipresent in computer systems. In this paper, we studied the adequacy of this metaphor in the context of a secure email system. Participant input suggests that representing certain privacy states, such as those concerning confidentiality and entity authentication, can be challenging and potentially problematic as a result of indicator misinterpretation. The simultaneous yet contradictory signals that can be communicated by an indicator, such as *security* and *cautiousness* in our use case, can impact the perceived security, or potentially the adoption of a system. While displaying an indicator set (e.g. via onboarding screens) could serve as a cue to engage users' familiarity with a specific concept and potentially prime users towards a specific goal, its effectiveness can be overshadowed by one or more contradictory indicators that constitute that set.

While further investigations within the context of use would be needed to validate the results, our findings highlight a larger problem. This goes beyond the simple design of an indicator, and more importantly it concerns the amount of security information that system designers try to communicate to users via an indicator. Making difficult design choices with respect to user-facing challenges such as entity authentication are widespread across security and privacy critical systems, thus, investigating those with representatives of the target population is a practice we strongly encourage as early as possible in the design process of privacy-enhancing technologies.

ACKNOWLEDGMENTS

We would like to thank p≡p for their collaboration and feedback as well as the anonymous reviewers for their comments and suggestions. Authors are supported by the Luxembourg National Research Fund through grant PRIDE15/10621687/SPsquared, and the project pEp Security SA / SnT "Protocols for Privacy Security Analysis".

APPENDIX

The surveys and datasets with raw participant responses from the three studies can be downloaded from the following link: http://doi.org/10.5281/zenodo.4322893

REFERENCES

[1] R Abu-Salma, M A Sasse, J Bonneau, A Danilova, A Naiakshina, and M Smith. 2017. Obstacles to the Adoption of Secure Communication Tools. In *2017 IEEE Symposium on Security and Privacy (SP)*. 137–153.

[2] Alex Ainslie, Adrienne Porter Felt, Robert W. Reeder, Somas Thyagaraja, Helen Harris, Alan Bettes, Jeff Grimes, and Sunny Consolvo. 2015. Improving SSL Warnings. (2015), 2893–2902.

[3] Bonnie Brinton Anderson, C. Brock Kirwan, Jeffrey L. Jenkins, David Eargle, Seth Howard, and Anthony Vance. 2015. How polymorphic warnings reduce habituation in the brain-insights from an FMRI study. *Conference on Human Factors in Computing Systems - Proceedings* 2015-April (2015), 2883–2892.

[4] W. Ross Ashby. 1956. *An Introduction to Cybernetics*. Chapman & Hall, London. http://pcp.vub.ac.be/books/IntroCyb.pdf

[5] Erinn Atwater, Cecylia Bocovich, Urs Hengartner, Ed Lank, and Ian Goldberg. 2015. Leading Johnny to Water: Designing for Usability and Trust. USENIX Association, 69–88.

[6] Wei Bai, Doowon Kim, Moses Namara, Yichen Qian, Patrick Gage Kelley, and Michelle L Mazurek. 2016. An Inconvenient Trust: User Attitudes Toward Security and Usability Tradeoffs for Key-Directory Encryption Systems. *Symposium On Usable Privacy and Security (SOUPS)* Soups (2016), 113–130.

[7] W Bai, D Kim, M Namara, Y Qian, P G Kelley, and M L Mazurek. 2017. Balancing Security and Usability in Encrypted Email. *IEEE Internet Computing* 21, 3 (2017), 30–38.

[8] Victoria Bellotti and Abigail Sellen. 1993. Design for Privacy in Ubiquitous Computing Environments. In *Proceedings of the Third Conference on European Conference on Computer-Supported Cooperative Work (ECSCW'93)*. Kluwer Academic Publishers, USA, 77–92.

[9] Jim Blythe, Jean Camp, and Vaibhav Garg. 2011. Targeted risk communication for computer security. *International Conference on Intelligent User Interfaces, Proceedings IUI* (2011), 295–298.

[10] Cristian Bravo-Lillo, Lorrie Faith Cranor, Saranga Komanduri, Julie Downs, and Saranga Komanduri. 2011. Bridging the Gap in Computer Security Warnings: A Mental Model Approach. *IEEE Security and Privacy* 9, 2 (mar 2011), 18–26.

[11] L. Jean Camp. 2011. Mental Models of Privacy and Security. *SSRN Electronic Journal* (2011).

[12] Lorrie Faith Cranor. 2008. A framework for reasoning about the human in the loop. *Proceedings of the 1st Conference on Usability, Psychology, and Security (UPSEC'08)* (2008), 1–15.

[13] Rachna Dhamija, J. D. Tygar, and Marti Hearst. 2006. Why Phishing Works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 581–590.

[14] Julie S. Downs, Mandy B. Holbrook, and Lorrie Faith Cranor. 2006. Decision strategies and susceptibility to phishing. *ACM International Conference Proceeding Series* 149 (2006), 79–90.

[15] V Dukhovni. 2014. *Opportunistic Security: Some Protection Most of the Time*. RFC 7435. RFC Editor. https://www.rfc-editor.org/info/rfc7435

[16] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. 2008. You'Ve Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 1065–1074.

[17] Adrienne Porter Felt, Robert W Reeder, Alex Ainslie, Helen Harris, Max Walker, Christopher Thompson, Mustafa Emre Acer, Elisabeth Morant, Sunny Consolvo, and U C Berkeley. 2016. Rethinking Connection Security Indicators. *the Symposium On Usable Privacy and Security (SOUPS)* Soups (2016), 1–14.

[18] Simson L Garfinkel and Robert C Miller. 2005. Johnny 2: a user test of key continuity management with S/MIME and Outlook Express. *Proceedings of the 2005 symposium on Usable privacy and security* 6 (2005), 13–24.

[19] Shirley Gaw, Edward W. Felten, and Patricia Fernandez-Kelly. 2006. Secrecy, flagging, and paranoia. (2006), 591.

[20] Jeffrey L. Jenkins, Bonnie Brinton Anderson, Anthony Vance, C. Brock Kirwan, and David Eargle. 2016. More harm than good? How messages that interrupt can make us vulnerable. *Information Systems Research* 27, 4 (2016), 880–896.

[21] Patrick W. Jordan. 2000. Inclusive design: An holistic approach. *Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Association, 'Ergonomics for the New Millennium'* (2000), 917–920.

[22] Kat Krol, Matthew Moroz, and M. Angela Sasse. 2012. Don't work. Can't work? Why it's time to rethink security warnings. *7th International Conference on Risks and Security of Internet and Systems, CRiSIS 2012* (2012), 1–8.

[23] Kenneth R. Laughery and Michael S. Wogalter. 2006. Designing Effective Warnings. *Reviews of Human Factors and Ergonomics* 2, 1 (2006), 241–271.

[24] Joscha Lausch, Oliver Wiese, and Volker Roth. 2017. What is a Secure Email? *EuroUSEC 2017* (2017).

[25] A Lerner, E Zeng, and F Roesner. 2017. Confidante: Usable Encrypted Email: A Case Study with Lawyers and Journalists. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. 385–400.

[26] Danielle Lottridge, Mark Chignell, and Aleksandra Jovicic. 2011. Affective Interaction: Understanding, Evaluating, and Designing for Human Emotion. *Reviews of Human Factors and Ergonomics* 7, 1 (2011), 197–217.

[27] Hernâni Marques and Bernie Hoeneisen. 2019. pretty Easy privacy (pEp): Mapping of Privacy Rating. IETF Internet-Draft, https://tools.ietf.org/html/draft-marques-pep-rating-01, Accessed: 10 October 2020.

[28] Hernâni Marques and Bernie Hoeneisen. 2019. pretty Easy privacy (pEp): Privacy by Default . IETF Internet-Draft, https://tools.ietf.org/html/draft-birk-pep-03, Accessed: 10 October 2020.

[29] Wolter Pieters. 2011. Explanation and trust: What to tell the user in security and AI? *Ethics and Information Technology* 13, 1 (2011), 53–64.

[30] Karen Renaud, Melanie Volkamer, and Arne Renkema-Padmos. 2014. Why Doesn't Jane Protect Her Privacy?. In *Privacy Enhancing Technologies*, Emiliano De Cristofaro and Steven J Murdoch (Eds.). Springer International Publishing, Cham, 244–262.

[31] Volker Roth, Tobias Straub, and Kai Richter. 2005. Security and usability engineering with particular attention to electronic mail. *Int. J. Hum. Comput. Stud.* 63, 1-2 (2005), 51–73. https://doi.org/10.1016/j.ijhcs.2005.04.015

[32] Scott Ruoti, Jeff Andersen, Scott Heidbrink, Mark O'Neill, Elham Vaziripour, Justin Wu, Daniel Zappala, and Kent Seamons. 2016. "We'Re on the Same Page": A Usability Study of Secure Email Using Pairs of Novice Users *(CHI '16)*. ACM, 4298–4308.

[33] Scott Ruoti, Jeff Andersen, Travis Hendershot, Daniel Zappala, and Kent Seamons. 2016. Private Webmail 2.0: Simple and easy-to-use secure email. *UIST 2016 - Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (2016), 461–472. arXiv:1510.08435

[34] Scott Ruoti, Nathan Kim, Ben Burgon, Timothy van der Horst, and Kent Seamons. 2013. Confused Johnny: When Automatic Encryption Leads to Confusion and Mistakes *(SOUPS '13)*. ACM, 5:1–5:12.

[35] Scott Ruoti and Kent Seamons. 2019. Johnny's Journey Toward Usable Secure Email. *IEEE Security and Privacy* 17, 6 (2019), 72–76.

[36] Itzel Vázquez Sandoval and Gabriele Lenzini. 2019. A Formal Security Analysis of the pep Authentication Protocol for Decentralized Key Distribution and End-to-End Encrypted Email. In *2nd International Workshop on Emerging Technologies for Authorization and Authentication, ESORICS International Workshops*.

[37] Jeff Sauro and James R. Lewis. 2016. *Quantifying the User Experience, Second Edition: Practical Statistics for User Research* (2nd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[38] Stuart E. Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. 2007. The Emperor's New Security Indicators. In *2007 IEEE Symposium on Security and Privacy (SP '07)*. 51–65.

[39] Steve Sheng, Levi Broderick, Jeremy J Hyland, and Colleen Alison Koranda. 2006. Why Johnny still can't encrypt: evaluating the usability of email encryption software. *Symposium On Usable Privacy and Security* (2006), 3–4.

[40] Borce Stojkovski and Gabriele Lenzini. 2020. Evaluating ambiguity of privacy indicators in a secure email app. In *Proceedings of the Fourth Italian Conference on Cyber Security, Ancona, Italy, February 4th to 7th, 2020 (CEUR Workshop Proceedings)*, Michele Loreti and Luca Spalazzi (Eds.), Vol. 2597. CEUR-WS.org, 223–234.

[41] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. 2009. Crying Wolf: An Empirical Study of SSL Warning Effectivenes. *18th USENIX Security Symposium* (2009), 399–432.

[42] Joshua Tan, Lujo Bauer, Joseph Bonneau, Lorrie Faith Cranor, Jeremy Thomas, and Blase Ur. 2017. Can Unicorns Help Users Compare Crypto Key Fingerprints?. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3787–3798.

[43] W. Tong, Gold S., S. Gichohi, M. Roman, and J. Frankle. 2014. Why King George III Can Encrypt. https://www.cs.princeton.edu/~arvindn/teaching/spring-2014-privacy-technologies/king-george-iii-encrypt.pdf

[44] Tara Whalen and Kori M. Inkpen. 2005. Gathering evidence: Use of visual security cues in web browsers. *Proceedings - Graphics Interface* (2005), 137–144.

[45] Alma Whitten and J D Tygar. 1999. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0 *(SSYM'99)*. USENIX Association, 14.

[46] Michael S Wogalter, David M DeJoy, and Kenneth R Laughery. 1999. Warnings and risk communication.

[47] Michael S. Wogalter and Kenneth R. Laughery. 1996. Warning! Sign and label effectiveness. *Current Directions in Psychological Science* 5, 2 (1996), 33–37.

[48] Min Wu, Robert C. Miller, and Simson L. Garfinkel. 2006. Do security toolbars actually prevent phishing attacks? *Conference on Human Factors in Computing Systems - Proceedings* 1 (2006), 601–610.