

“Something isn’t secure, but I’m not sure how that translates into a problem”:

Promoting autonomy by designing for understanding in Signal

Justin Wu
Brigham Young University

Cyrus Gatrell
Brigham Young University

Devon Howard
Brigham Young University

Jake Tyler
Brigham Young University

Elham Vaziripour
Utah Valley University

Kent Seamons
Brigham Young University

Daniel Zappala
Brigham Young University

Abstract

Security designs that presume enacting secure behaviors to be beneficial in all circumstances discount the impact of response cost on users’ lives and assume that all data is equally worth protecting. However, this has the effect of reducing user autonomy by diminishing the role personal values and priorities play in the decision-making process. In this study, we demonstrate an alternative approach that emphasizes users’ comprehension over compliance, with the goal of helping users to make more informed decisions regarding their own security. To this end, we conducted a three-phase redesign of the warning notifications surrounding the authentication ceremony in Signal. Our results show how improved comprehension can be achieved while still promoting favorable privacy outcomes among users. Our experience reaffirms existing arguments that users should be empowered to make personal trade-offs between perceived risk and response cost. We also find that system trust is a major factor in users’ interpretation of system determinations of risk, and that properly communicating risk requires an understanding of user perceptions of the larger security ecosystem in whole.

1 Introduction

The primary goal of usable security and privacy is to empower users to keep themselves safe from threats to their security or privacy. Their ability to do so is reliant on an accurate assessment of the existence and severity of a given risk, the set of available responses, and the cost of enacting those responses. Ideally, users would like to take action only when

a threat has been realized *and* the negative consequences of that threat are severe enough to outweigh the costs of enacting the mitigating measure. In practice, however, it is difficult for users to have a comprehensive view of the situation and thus make informed decisions. Typically developers of secure systems best understand the nature of risks users will encounter and design responses that will mitigate those risks, but it is difficult for them to communicate this knowledge to users who are ultimately responsible for weighing risk severity and response cost trade-offs.

Consequently, the design of many security mechanisms seeks to simplify the threat-mitigation equation by avoiding calculations of risk impact and response cost, either through automating security measures or by pushing users to unilaterally enact protective measures regardless of context. This approach, however, is not without drawbacks. It discounts the impact of response costs on users’ lives by presupposing that the execution of a protective behavior is always a favorable cost-benefit proposition. In reality, however, the “appetite and acceptability of a risk depends on [users’] priorities and values” [12]. Indeed, it has been argued that, “Security that routinely diverts the attention and disrupts the activities of users in pursuit of these goals is thus the antithesis of a user-centered approach” [20].

This approach and its drawbacks is evident in the current design of secure messaging applications. In a typical secure messaging application, an application server registers each user and stores their public key. When a user wishes to send a secure message to someone, the application transparently retrieves the public key of the recipient from the server and uses it to automatically encrypt messages. However, because the server could deceive the user, either willingly or because it has been coerced by a government or hacked by an attacker, communicating parties must verify one another’s public keys in order to preserve the cryptographic guarantees offered by end-to-end-encryption. The method by which parties verify their public keys has been called the authentication ceremony, and typically involves scanning a contact’s QR code or making a phone call to manually compare key fingerprints.

The usability of the authentication ceremony in secure messaging applications has been studied in recent years, with the general conclusion that users are vulnerable to attacks, struggling to locate or perform the authentication ceremony without sufficient instruction [1, 21, 28]. The root cause of this difficulty is that the designers of these applications do not effectively communicate risks, responses, and costs to users. The automatic encryption “just works” when there is no attack, but the application does not give users enough help to judge risk and response trade-offs when an attack is possible. Prior work [29] applied opinionated design to the Signal authentication ceremony and showed that they could significantly decrease the time to find and perform the authentication ceremony, with strong adherence gains. However, this work assumed that all users should perform the ceremony for every conversation, when many users may not want to incur this cost due to low perceived risk or high response cost.

In this study, we demonstrate an alternative design approach that emphasizes users’ comprehension over compliance, with a goal of empowering users to make more informed decisions that align with their personal values. We employ a design philosophy that might be seen as partway between opinionated and non-opinionated design: our design pushes users to make decisions, but not any decision in particular.

To this end, we conduct a three-phase redesign of the warning notifications surrounding the authentication ceremony in the Signal secure messaging app. We use Signal because the Signal protocol has been the foundation upon which other secure messaging applications have built, and thus many secure messaging applications share its basic design features and have similar authentication ceremonies. Because Signal is open source, we can apply design changes and, if these changes are successful, influence applications based on Signal, such as WhatsApp and Facebook Messenger.

The authentication ceremony in Signal is a particularly good fit for applying a risk communication approach to design. First, the system has an explicit and timely heuristic for identifying shifts in risk levels: encryption key changes. Moreover, because changes in security state are contingent upon key changes, we need only communicate with users once a potential risk occurs. Furthermore, the available mitigating response to a key change is unambiguous: performing the authentication ceremony. Finally, the authentication ceremony is a mechanism where response cost factors heavily into the equation—users must be synchronously available to perform it—even though most key changes are due to reinstalling the application, not a man-in-the-middle attack.

Our redesign generally follows a standard user-centered design process, but with an explicit focus on enabling users to make more informed decisions. First, we measured the baseline effectiveness of Signal’s man-in-the-middle warning notifications with a cognitive walkthrough and a lab-based user study. Next, we designed a set of candidate improve-

ments and evaluated their effectiveness by having participants on Amazon’s Mechanical Turk platform interact with and rate design mockups. Lastly, we implemented selected improvements into the Signal app and evaluated our redesign with a user study that repeated the conditions of the first study.

We make the following contributions:

- **Identify obstacles to user understanding of the authentication ceremony in Signal.** We performed a cognitive walkthrough of Signal’s authentication ceremony and associated notifications, highlighting barriers to understanding its purpose and implications. We followed up on our findings with a user study exposing participants to a simulated attack scenario, which allowed us to evaluate the effectiveness of these warnings in practice.
- **Perform a comprehension-focused redesign of the authentication ceremony with an aim at empowering users to balance risk-response trade-offs in a manner concordant with their personal priorities.** Building on the findings of our cognitive walkthrough and user study, we redesigned the authentication ceremony and associated messaging with a focus on empowering users to make more informed decisions. Candidate designs were evaluated by users on Amazon Mechanical Turk with a final redesign evaluated in a user study. Our redesign results in higher rates of both comprehension and adherence as compared to Signal’s default design.
- **Show that risk communication empowers users to decide that not enacting protective behaviors is the right choice for them.** We find evidence that making users aware of the presence of an active threat to their data privacy is insufficient to produce secure behaviors. Users instead weigh the perceived impact of negative outcomes against the cost of enacting the response. Because “worst-case harm and actual harm are not the same” [10], this balancing of trade-offs can weigh unfavorably against performing protective measures.
- **Show that users’ strategies for mitigating perceived threats are dependent on their perception of the larger security ecosystem as a whole.** Despite our redesign prompting a greater share of users to perform the authentication ceremony, and producing greater understanding of the purpose thereof, participants’ preferred strategies for mitigating the perceived interception risk did not change substantially. Instead, it is apparent that users have developed an array of protective behaviors they rely upon to ensure positive security and privacy outcomes that exist beyond the ecosystem of any given app or system.

Artifacts: A companion website at <https://signal.internet.byu.edu> provides study materials, source code, and anonymized data.

2 Related work

2.1 Protection motivation theory

We base our work on protection motivation theory (PMT), which tries to explain the cognitive process that humans use to change their behavior when faced with a threat [14, 19]. The theory posits that humans assess the likelihood and severity of a potential threat, appraise the efficacy and cost of a proposed action that can counter the threat, and consider their own efficacy in being able to carry out that action.

Recently, PMT has been applied to a variety of security behaviors. Much of the work in this area is limited to studying the *intention* of individuals to adopt security practices, such as the intention to install or update antivirus software, a firewall, or use strong passwords [13, 32]. However, psychological research has demonstrated there is a gap between intention and behavior [22, 23], similar to the gap reported between self-reported security behaviors and practice [30]. A few studies have used objective measures of security behavior to study connections to PMT, such as compliance with corporate security policies [32], adoption of home wireless security [31], and secure navigation of an e-commerce website [27].

2.2 Risk communication

We are interested in studying how application design can be modified to help users assess risk and thus make more informed choices. We thus draw upon the wide variety of work in usable security that has focused on the design of warnings given to users.

Microsoft developed the NEAT guidelines for security warnings [18], emphasizing that warnings should only be used when absolutely *necessary*, should *explain* the decision the user needs to make, should be *actionable*, and should be *tested* before being deployed. Browser security warnings, in particular, have had a long history of lessons learned, including eliminating warnings in benign situations [26], removing confusing terms [4], and following the NEAT guidelines [8]. Phishing warnings are recommended to interrupt the primary task and provide clear choices [6]. Other work has recommended that software present security behaviors as a gain and use a positive affect to avoid undue anxiety [9].

We also draw upon risk communication, a discipline focused on meeting the need of governments to communicate with citizens regarding public health and safety concerns [5]. Nurse et al. provide a summary of how risk communication can be applied to online security risks [16]. Their recommendations include focusing on reducing the cognitive effort by individuals, presenting clear and consistent directions for action, and presenting messages as close as possible to the risk situation or attack. One noteworthy effort used a risk communication framework to redesign warnings for firewall software [17]. Their results show that the warnings improved com-

prehension and better communicated risk and consequences. However, the focus of this study, as with many others, was on greater compliance with recommended safe behaviors.

In contrast, we feel that risk communication provides a greater benefit in usable security when it enables users to make rational decisions based on their values, as opposed to compliance with a prescriptive behavior that experts believe is correct. For example, Herley has emphasized the rationality of users' rejection of security advice, by explaining that users understand risks better than security experts, that worst-case harm is not the same as actual harm, and that user effort is not free [10]. Sasse has likewise warned against scaring or bullying people into doing the "right" thing [20]. Indeed, recent work on what motivates users to follow (or not follow) computer security advice indicates that differences in behavior stem from differences in perceptions of risk, benefits, and costs [7].

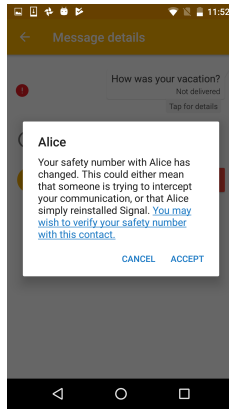
As stated by the National Academies, "*citizens are well informed with regard to personal choices if they have enough understanding to identify those courses of action in their personal lives that provide the greatest protection for what they value at the least cost in terms of those values*" [5]. Success is measured in terms of the information available to decision makers, and need not result in consensus or uniform behavior due to differences in what individuals value or perceive in terms of risks or costs of action.

3 Evaluating warnings in Signal

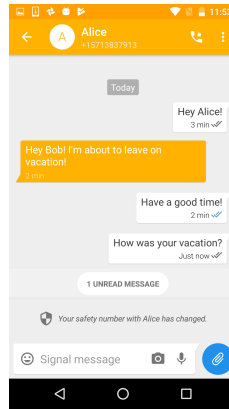
Signal uses the phrase *safety number* to describe a numeric representation of the key fingerprints for each participant in a conversation, warning users when this safety number changes. A safety number change occurs either when someone reinstalls the app (which generates new keys), or if a man-in-the-middle attack is conducted, with an attacker substituting their own key for an existing one. The authentication ceremony in Signal is referred to as *verifying* safety numbers; matching safety numbers rules out an attack. To evaluate the effectiveness of notifications that Signal currently uses we conducted both a cognitive walkthrough and a lab user study.

3.1 Cognitive walkthrough

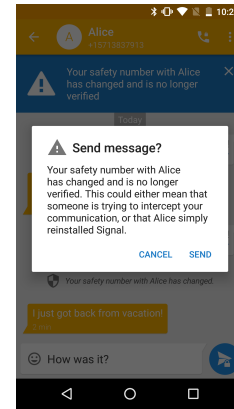
We performed a cognitive walkthrough of the notifications presented to users when a key change occurs and the authentication ceremony. The walkthrough was conducted by four of the authors, with a range of experience—a professor and a graduate student with substantial prior HCI and Signal research and two undergraduate students with no prior experience with HCI or with Signal. Our walkthrough consisted of exposing the user to every possible scenario leading to a safety number change, documenting all notifications and messages that are presented to the user and mapping the flow of decisions the user can make at each point. In addition, we



(a) Message not delivered dialog



(b) Shield message



(c) Message blocked dialog

Figure 1: Signal notifications when safety numbers differ, depending on the internal state of the application.

analyzed Signal’s code base to establish how internal state accompanied each warning notification and the effects of user actions on these states.

Our cognitive walkthrough revealed that, depending on the internal state of the system prior to a key change, Signal will react in one of three different ways to a key change event, as depicted in Figure 4 in Appendix A:

- *Message not delivered (top path in Figure 4):* This path is activated when the user has not previously verified safety numbers, is still on the conversation screen, and attempts to send a message. Sent messages will show up in the conversation log, accompanied by a notification informing the user that they were “not delivered” and that they may tap for more details. Doing so brings up another screen which clarifies that there is a “new safety number” alongside a “view” button. Tapping the button generates a dialog (Figure 1a) with a succinct message about safety number changes and several options for proceeding, including one that leads to the authentication ceremony screen and one that clears the warning state.
- *Message delivered (bottom path in Figure 4):* This path is activated when the user has not previously verified safety numbers and has either left the conversation screen or received a message. Signal will insert a notification into the conversation log informing the user of a safety number change, using a shield icon to mark the notification (Figure 1b). Tapping this dialog will take the user to the authentication ceremony screen. The shield and message appear in all three flows, but this is the only notification given to users in this flow; no other changes occur.
- *Message blocked (middle path in Figure 4):* This path is activated when the user has previously verified safety numbers and has either left the conversation screen or received a message. This scenario places a blue banner at the top of the conversation log, warning users that their “safety

number has changed and is no longer verified”. Tapping this banner takes users to the authentication ceremony. If the user attempts to send a message while in this state, Signal will prevent the message from being sent, and a dialog will be shown (Figure 1c). This dialog informs users that the safety number has changed and asks whether they wish to send the message or not. The user has three ways to clear the warning state in this scenario. They may select the “send” option at the dialog, mark the contact as verified on the authentication ceremony screen, or tap the “x” on the blue banner.

Our cognitive walkthrough identified numerous issues that may be confusing and that contradict recommendations on effective warning design:

- *Unclear risk communication.* It may not be clear to users what the term “safety number” means, nor what it means that these have changed.
- *Inconsistency of choice across dialogs.* Although the message-not-delivered and message-blocked flows show dialogs that convey nearly identical messaging, they present users with different choices for interaction (Figures 1a and 1c respectively).
- *The consequences of user actions are not clear beforehand.* For example, in the message-not-delivered flow, it is likely for the user to send multiple messages that are blocked from delivery before noticing and attempting to resolve the error. If the user selects “Accept” at the ensuing dialog, this will automatically re-send *all* failed messages; not just the one selected for inspection. Conceivably, should one or more of those failed messages contain sensitive information, this might be undesirable behavior.
- *The implications of success or failure of the authentication are unclear.* In the event of a failed safety number

match—the identification of which is the entire reason for the authentication ceremony—no recommendations for subsequent action are made to the user.

- *Does not communicate response cost.* The costs and requirements for performing the authentication ceremony are not made clear before users are brought to the authentication ceremony screen.

3.2 User study #1: Methodology

The following study, and all others in this work, were approved by our Institutional Review Board.

We designed a between-subjects user study to evaluate the effectiveness of each of these three notification flows at informing users of the potential risks they face and the responses available to them when exposed to a man-in-the-middle attack scenario. To control environmental conditions all participants used a Huawei Mate SE Android phone that we supplied.

For each of the three notification flows we discovered in our cognitive walkthrough, 15 pairs of participants (for a total of 45 pairs) conducted two simple conversation tasks. A simulated man-in-the-middle attack was triggered between the first and second tasks, causing the corresponding warning notifications to appear for each participant at the start of their second task. We simulated the attack by modifying the Signal source code to contact a server we operate and then change the encryption keys on demand. Participant reactions were recorded with video and a post-task questionnaire.

Our choice of tasks differs from previous work that asked participants to transmit sensitive information. Instead, we had participants communicate non-sensitive information, because this has the potential to reveal more diverse behaviors when faced with a risk of interception. For example, some users may be unconcerned by interception or unwilling to incur the cost of conducting the authentication ceremony if they perceive a conversation with non-sensitive information to be low risk. Others, on the other hand, may still find a potential attack to be unsettling and thus assess the risk to be more severe and/or the cost to be more worthwhile. A scenario with sensitive information could interfere with this dynamic.

We performed the studies for each treatment type—each notification flow—in succession, such that the first 15 pairs all experienced the message-not-delivered flow, the next 15 pairs saw only the message-delivered flow, and the final 15 pairs were exposed to the message-blocked flow.

3.2.1 Recruitment and Demographics

We recruited participants by posting flyers in buildings on our university campus. The flyer instructed participants to bring a partner to the study. Participants were each compensated \$15, for a total of \$30 per pair. Studies lasted approximately 40 minutes.

Our sample population skewed young, with 92.2% ($n=83$) of our participants aged between 18-24. Our population also skewed female (61.1%, $n=55$). A skills-based, self-reported assessment of technical familiarity revealed a normal distribution with most participants familiar with using technology.

3.2.2 Study design

When participants arrived, they were randomly assigned to an A or B roleplay condition (with a coin flip). Participants were then escorted to separate rooms, where they were presented with a packet of instructions, with one page per task.

Participants were first directed to register the Signal app pre-installed on the phones, granting all permissions the app sought in the process. Once both participants had finished registration, they were directed to begin their first task: to coordinate a lunch appointment using Signal. This task was designed to familiarize our participants with the operation of Signal. Exchanging messages is also necessary for Signal to establish safety numbers that could then be changed as part of the man-in-the-middle-scenario.

Next, participant B's roleplay informed them that participant A had gone to Hawaii on vacation, and to hand their phone to their study coordinator to simulate this communication disconnect. Participant A's roleplay provided similar information, including the instruction to hand their phone to their study coordinator, but additionally provided a half-page description of their "trip".

Study coordinators took this opportunity to manipulate Signal into the conditions necessary for the associated treatment as well as triggering the simulated man-in-the-middle attack. Finally, phones were handed back to participants, and they were instructed to continue on to their final task.

Finally, participants were instructed to discuss and share photos of participant A's trip to Hawaii, which had been preloaded onto participant A's phone. With the simulated attack active, participants were now exposed to the warning notifications corresponding to their treatment group. These final instructions explicitly stated that participants were finished with this task whenever *they* believed they were, to avoid biasing participants toward any particular action in the event of a failed authentication ceremony.

Once both participants declared the task complete, they were given the post-task questionnaire. This questionnaire asked them if, within the context of their roleplay, they had perceived a risk to their privacy. They were then asked how they might mitigate this risk, and to describe how effective they believe their strategy would be. Finally, participants were shown each of the warning notification elements in turn, and asked: (1) whether or not they had seen them, (2) what message they believed the notification was attempting to convey, and (3) what effects they believed the associated interactive elements would produce.

Upon completion of the questionnaire, participants were read a short debrief, informing them that the attack had only been simulated, that Signal employs multiple features intended to both prevent and identify interception, and that no such attacks have ever been reported in the wild.

3.2.3 Data analysis

All open-ended questionnaire responses were coded by two of the authors in joint coding sessions using a conventional content analysis approach [11].

3.3 User study #1: results

3.3.1 Risk perception and mitigation

Roughly half of groups 1 and 3, the treatment groups whose messages either failed to send or were blocked, perceived a risk during the study scenario (13/30 and 16/30 participants respectively). In stark contrast, however, only a small fraction of the participants in group 2 (4/30), whose workflow was not interrupted, felt that they had encountered a risk. In explaining the nature and properties of the risk they perceived, participant responses generally fell in one of three categories: (1) a security risk of an unknown nature, (2) a risk of interception, or (3) a risk of an insecure communication channel. Perceptions of how to mitigate such a risk generally fell under one of three categories: self-filtering (avoiding communicating sensitive information), use of an alternative communication channel such as another app, and verifying a contact.

3.3.2 Shield message

The shield message in the conversation log, “Your safety number with <contact> has changed”, confused a number of participants. While many participants correctly associated this message with a change in security status, a number interpreted it to mean precisely the opposite of its actual meaning—that it conveyed *improved* security levels. As one participant explained following our post-study debrief, “*I thought that it was improving security—that every once in a while, you change the safety number so it refreshes and makes it harder for people to hack into. So, I was like, ‘Oh, it’s doing its job.’ Apparently, it wasn’t!*”

Next, as our cognitive walkthrough predicted, participants were confused by what, precisely, it was that had changed, offering numerous different explanations. Examples include: phone number, connection, safety number, safety code, “something technical”, settings, security code, and verification code. As one participant remarked, “*Some sort of safety code changed. Or his actual phone number, I was a little confused.*”

Participants acted on this message all cited the importance of ensuring privacy/security outcomes. Those who did not act on it did so because: (1) they did not see it as an actionable message, (2) they explicitly expressed having been habituated

against such notifications, (3) the information they were communicating was seen as non-sensitive, or (4) they perceived it to be a part of the study task.

Notably, perceptions of the non-sensitivity of the conversation were critical in putting participants at ease even if they had found the notification alarming, as exemplified by one participant response: “*I felt that it was important because of the nature of the app and whenever a safety anything is changed that usually is noteworthy. I would have put that it was extremely important if I had felt like there was an actual risk of someone actually trying to read our conversation.*”

3.3.3 Message-not-delivered dialog

Only participants in treatment group 1 were exposed to the message-not-delivered dialog. Participants were asked to describe what they believed would happen if they were to tap the three interactive elements in this dialog: the “Accept” and “Cancel” buttons and the link embedded in the text.

Participants generally understood that “Cancel” would leave the system state unchanged. Similarly, most participants understood that “Accept” would unblock their messages and allow them to communicate once more. Perception of the link, however was more confused. 9 of the 14 participants who responded to this question responded that they had believed it would have taken them to a screen explaining more about the situation. This is in contrast to what it really does, which is to redirect users to the authentication ceremony, as noted by one participant who expected it to lead “*to an ‘About’ or ‘Info’ page, but it ended up taking me to the verification.*”

3.3.4 Blue banner & message-blocked dialog

Understanding of the options presented by the message-blocked dialog—“Send” and “Cancel”—were high. However, unlike the message-not-delivered dialog, the message-blocked dialog does not present a method to reach the authentication ceremony—instead accessible via the blue banner.

Understanding of the blue banner was mixed among those participants of group 3 who reported having seen it. Only roughly half understood that it was a privacy-related warning. Others were either entirely at a loss to explain its purpose or believed that it was a system error notification. Those who were confused by its meaning or believed it to be a system error did not feel it warranted action. Of the five participants who correctly interpreted the blue banner as a warning, two did not feel they were at risk, and thus did not feel like action was warranted.

3.3.5 Authentication ceremony

Participants who reported having seen the authentication ceremony screen were asked about the significance of verifying safety numbers (and whether or not they matched) as well as about the verification toggle. Participants may have seen,

and even interacted with, the authentication ceremony screen without necessarily having performed the authentication ceremony. In total, 5 pairs of participants conducted the authentication ceremony while 27 participants reported having seen the screen.

As predicted in our cognitive walkthrough, participants were confused about what a safety number was or why it had changed. For instance, one participant explained that *“I honestly wasn’t sure what it meant. I didn’t know that I had a safety number with them in the first place so I was unaware that it could change.”* We also noted occasions where participants entered the authentication ceremony screen only to back out without completing it. This may be due to poor communication regarding response cost—both conversation partners should either be in the same physical location to execute the QR-code ceremony or be willing to verify safety numbers over another medium (such as a phone call).

Also as predicted, the verification toggle confused participants. Of the 11 participants who reported having flipped the toggle, not one participant correctly intuited its use. 7 of these 11 toggled it purely as an exploratory action, unaware that doing so would inadvertently and incorrectly clear the warning state.

When asked to characterize the purpose of the authentication ceremony, participants did generally associate it with verification, although their model for *what* it verifies was often incorrect. Table 2 shows a qualitative analysis of participant responses when asked the purpose of the ceremony, and the meaning of a matching or non-matching result, with responses coded and then categorized as correct, partially correct, or incorrect. Only a few participants understood that the purpose of the authentication ceremony is to verify the confidentiality of the conversation. Instead, a number of participants mistakenly believed that it was about verifying the identity of the individual, i.e., that *“it makes sure the other person is who you think they are”*, as one participant explained. This threat model does not account for a different type of attacker the authentication ceremony is intended to detect: a passive man-in-the-middle who simply decrypts and forwards messages without interfering in the conversation.

These misconceptions naturally carried forward into responses about the significance of matching and non-matching safety numbers. Perceptions of non-matching safety numbers correctly assessed this result as indicative of interception occurring, but again, participants often believed that this meant that they had detected an impersonator, as with one participant who remarked that, *“Someone using another phone could be posing as my brother, I guess.”* Participants did almost unilaterally understand that matching safety numbers were indicative of a positive security/privacy outcome, although several participants misinterpreted the role of the authentication ceremony as a mechanism that would actively *prevent* interception, as opposed to detecting it.

4 Developing improvements

Based on the results of our cognitive walkthrough and subsequent user study, we concluded that there were three main areas for improvement worthy of focus: (1) the need for an accessible, persistent visual indicator for verification state, (2) the messaging used in warning notifications and dialogs, and (3) the notification flow and all associated UI elements.

4.1 Visual indicator

Visual indicators, or icons, are important both as an accessible measure for communicating security state to users with a single glance as well as for enhancing the consistency of warning notifications. While the authentication ceremony screen in the original version of Signal does have a (somewhat hidden) lasting representation of verification state, the verified toggle switch, we believe that this indicator is inadequate because it represents only two states (verified and unverified) and because it confused users in our lab study who believed that toggling the switch would verify their partner.













We decided to create a set of icons that would properly reflect all three verification states: (1) the default, assumed-safe state of the conversation prior to a safety number change, (2) a verified state that reflects matching key fingerprints, and (3) an unsafe state that reflects having found non-matching fingerprints in the authentication ceremony. Ideally, the icon for the default state could have a small modification to represent the other two states. By adding this visual indicator onto the action bar, it becomes both an accessible indicator of state as well as a shortcut to the authentication ceremony.

We began by designing a neutral icon to represent the default state. Our goal was to select an icon that would be intuitively associated with privacy, and that would not evoke unwarranted feelings of concern, since this state does *not* signal a cause for concern. We selected a blank shield icon for this purpose. We then created variants of this icon, as shown in Table 1 to represent the success and failure states post-authentication ceremony.

We evaluated our designs on Amazon’s Mechanical Turk platform, with each icon being shown to at least 50 participants. Each icon was shown occupying a position on the action bar in a screenshot of Signal’s interface, next to the call button. For positive-valenced icons we asked participants to rate how strongly they associated the icon with privacy on a scale from 1-10. For negative-valenced icons we asked participants to rate how worried they would feel if they saw the associated icon. We asked both questions for the blank shield icon.

As shown in Table 1, the blank shield has a moderate association with privacy and a low association with worry, making it a good fit for a default icon. We discounted any icons using a lock because it is used elsewhere in the app to represent encryption, and we wished to avoid conflating meanings. We

Table 1: Comparison of icons a 10-point Likert scale

Icon	Mean	Std. Dev.	Count
<i>Positive – association with privacy</i>			
	6.50	2.21	74
	6.54	2.82	79
	5.74	2.56	78
	7.52	2.43	83
<i>Negative – association with worry</i>			
	4.08	2.36	59
	4.95	2.36	55
	5.11	2.64	57
	4.17	2.51	65
	4.95	2.49	60
	4.52	2.53	52
	5.56	2.53	61
	5.05	2.51	62

chose the shield with a checkmark enclosed by a circle for the positive icon because of the remaining choices it had the strongest association with privacy. Surprisingly, no negative icon evoked strongly negative associations. We chose the shield with an exclamation mark because it had the strongest negative associations, and if the privacy check fails we do want users to be alarmed.

Appendix C shows how these indicators are used in our design.

4.2 Notification and dialogs

We revised notifications and dialogs concerning safety numbers throughout Signal by following recommendations for warning design and risk communication. The principles we followed are (a) interrupt the primary task, (b) present messages close to the risk situation, (c) reduce cognitive effort (d) use a positive affect, (e), explain the decision the user needs to make, and (f) present clear and consistent directions for action. In particular, we designed the following changes, with screenshots shown in Appendix B:

- *Positive framing for the authentication ceremony.* We framed the authentication ceremony as a “privacy check”, which emphasizes the *role* it plays rather than the primitives or actions involved, which will be unfamiliar to users. Notifications of changed safety numbers (what we refer to as the “shield message”) instead report that Signal recom-

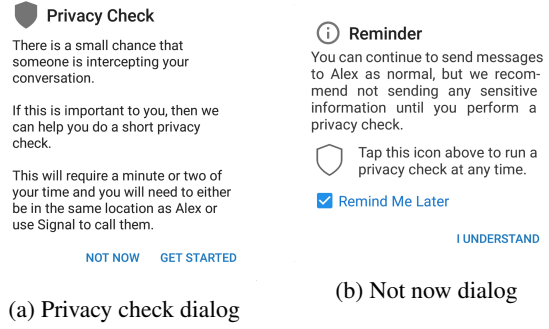


Figure 2: New notification dialogs, framing the authentication ceremony as a privacy check and using risk communication principles.

mends a privacy check, turning what was sometimes seen as a routine system notification into an explicitly actionable recommendation. We also frame the consequences of performing the privacy check in a positive manner such that both positive and negative results present benefits for the user: a positive match guarantees conversation privacy and a failed match reveals ongoing message interception. In this way, even if fingerprints do end up matching—by far the most common case—users need not feel that it was a waste of their time to engage in the verification process.

- *Communicating response cost and providing users with alternatives to the privacy check.* Our dialogs inform participants up-front what executing the privacy check will require (Figure 2a), with a “Not now” option that generates a reminder dialog for participants who are uninterested (Figure 2b). The reminder dialog includes a recommendation to not communicate sensitive information until the privacy check is first completed, with an option be reminded at a later time and a description of how to access this functionality at any time. Thus, participants’ options are framed as clear choices with defined costs.
- *Safety number labeling and interaction changed.* To promote better understanding of the safety numbers and their role, we divided the safety numbers into their constituent halves, relabeled them as *device identifiers*, and explained that they are used in encrypting the conversation.¹ Prior work indicated users dislike how long the safety number is [28]. Thus, we also rearranged groupings from 5 digits in a set to 3. This aligns more naturally with the standard process for grouping numbers, where numbers larger than 999 are grouped into sets of three known as periods. This does not reduce the actual count of numbers, but does reduce cognitive load.

¹This is not technically accurate, as they are key fingerprints and not keys themselves, but our goal is to have participants associate the comparison task with the preservation of a secure conversation and not to overwhelm them with details of the encryption process.

Privacy Verified!

Congratulations! You and Alex have checked your device identifiers and found that they match.

This means no one, not even Signal, can read or listen to your conversation. We will notify you if you ever need to check again.



Alex has been marked as verified

BACK TO PHONE CALL

(a) Success dialog

Warning

You have found that your identifiers with Alex do not match. You should double check that they do not match.

If you're sure you compared Alex's identifier, this means that someone is likely able to read and listen to your conversation. We recommend that you do not communicate any sensitive information.

☒ Send a report to Signal

TRY AGAIN IM SURE

(b) Failure dialog

Figure 3: New authentication ceremony success and failure dialogs.

We provide two options for performing the authentication ceremony, an in-person QR-code scan and a phone call comparison, as recommended in [29]. In the phone call version, we removed the confusing toggle element, which we replaced with two buttons explicitly labeled “Match” and “No match”.

- *Addition of success and failure messaging when the privacy check is completed.* We added dialogs after the privacy check that inform users of the implications of success and failure (Figures 3a and 3b respectively). We also designed a dialog shown before the authentication ceremony. If the privacy check has already been completed, it will show the current state and its implications for the user; if it has not, then it will instead explain what the privacy check entails and provide access to our authentication ceremony options.
- *Options for interaction inform users of the choice they're making.* To promote user autonomy and informed decisions, we carefully selected labels for our dialog buttons that describe the consequences of that choice and imply active decision-making on the part of the user, such as “Not now” and “Get started” on the privacy check dialog, as opposed to the more traditional “Okay” and “Cancel”.

4.3 Notification flow

As described earlier, based on the system state prior to a key change, Signal diverges into one of three different notification flows. In order to provide a consistent user experience, we decided to instead use a single, unified flow every time a key change occurs. We eliminated the non-interrupting flow from consideration because in our first study it was ineffective at promoting either adherence or comprehension. This left us with the two interrupting flows, the message-not-delivered and message-blocked flows, which produced similar comprehension levels in our user study. We hypothesized there might be a difference between these because the timing of interruptions can have an impact on decision-making [2, 3, 15, 25]. We further identified two additional UI elements that might

contribute to our aims of increased comprehension: (1) an introduction screen showing the privacy check icons after registration and (2) the blue banner element that accompanies the message-blocked flow in the original Signal.

To evaluate the relative effectiveness of these elements we designed a website containing a simulated Signal experience using mockups of our candidate flows and had users of Amazon’s Mechanical Turk platform interact with the simulation using a between-subjects comparison. Participants were, as in our user study, presented with two simple communication tasks involving non-sensitive information and a man-in-the-middle occurring between the first and second task. Unlike in our user study, users selected from a set of predefined messages, although they otherwise interacted with the interface normally, and their interactions were recorded in a database. For example, participants that wished to proceed with the privacy check were shown the results of the check as if they performed the authentication ceremony and asked to choose a response from the resulting dialog. After they were done, participants were given a tailored questionnaire which asked their perception of the notifications they had seen as well as why they had chosen the options they did. A total of 223 participants interacted with mockups and explained their actions via a post-task questionnaire.

We separated the elements to be evaluated into three rounds. The first round compared our delivery mechanisms: the message-not-delivered and message-blocked flows. The winner of the first round was then evaluated against a version that also included the blue banner element. Finally, the winner of the second round was then evaluated against a version that added an introductory screen.

To test for the difference between the message-not-delivered and the message-blocked flows we measured how many participants chose to start the authentication ceremony. We observed no significant difference (35/50 vs 31/50). We opted to use the message-blocked control flow because the message-not-delivered flow complicates the user’s task when they must resolve multiple failed messages. To test whether the blue banner message had an improvement we again measured how many participants chose to start the authentication ceremony and observed no difference (31/50).

To test whether the introductory screen had a difference we qualitatively measured comprehension. To do this we used participant responses to a question asking them what the privacy check notification meant. Several authors coded each response and then determined whether the participant understood that this notification meant interception of their conversation could be happening and found a slight improvement with the introductory screen (30/50 vs 23/50). However, we chose to leave the introductory screen out of our final design because the effect was not large and could have been exaggerated due to the short-term nature of the simulation.

Qualitative analysis of participant responses regarding their decision to perform (or not perform) the privacy check showed

participants weighed risks with response costs and made reasoned choices. Roughly 60% of all groups opted to perform the privacy check, with the remainder choosing the other option, “not now”. Participants who opted to perform the privacy check typically stated having done so out of a desire to verify the existence of the risk, because they believed it better to be safe than sorry, or out of curiosity. Participants who chose “not now” had either determined the risk to be of minimal severity or because they felt executing the privacy check would be inconvenient. Those who felt it would be inconvenient described it as such either because the current timing was seen as inappropriate or because of the synchronization cost (needing both members of the pair to execute the privacy check at the same time).

5 Evaluating the effectiveness of our redesign

We conducted a lab study to evaluate the effectiveness of our changes. Appendix B shows the control flow we used and screenshots of the new notifications and UI elements, and Appendix C shows the new indicators. We maintained the same study design used in our first lab study, with some minor modifications. Since our redesign has just one control path, we ran this study with just one treatment group of 15 pairs. Because we included additional screens that have no analogous equivalent in the original version of Signal, the post-task questionnaire in this study is not fully comparable with that from our first.

5.1 Results

5.1.1 Risk perception

Two-thirds (20/30) of our final user study participants reported having perceived a threat within the context of their roleplay. Qualitative responses indicated participants largely correctly perceived an interception risk, while a handful, interestingly, believed that Signal was itself the risk; this view seemed to be fueled by the number of permissions that Signal asks for in short succession. Participant notions of how they might mitigate perceived risks virtually mirrored those from the first user study—self-filtering, using alternate communication channels, and verifying contacts—along with restricting app permissions. Notably, only a small fraction of open responses (2/20) mentioned the privacy check as their mitigating strategy of choice despite, as we describe shortly, improved adherence and comprehension rates.

5.1.2 “Signal recommends a privacy check”

Qualitative responses indicate nearly all participants associated this notification with security, although as with our first study, there were a few who misinterpreted it as an *increase* in security. Due to our removal of Signal’s original messaging

regarding a change, participants of our final user study were not confused about what had changed as the first groups had been.

Those who felt it important to act upon this message generally explained that they felt ensuring privacy outcomes to be important, as with one participant who explained, “*I hear a lot about data breaches and such, so seeing that the app was giving a warning notification showed to me that it was something important that I should act on.*” Importantly, those who did not feel that the notification was cause for concern typically felt that way because the information they were communicating was perceived as non-sensitive in nature.

5.1.3 Privacy check dialog

Qualitative responses indicate participants generally understood the dialog was informing them of a potential threat, although perceptions of the nature of that threat and of the likelihood of that threat were more varied. For example, while most participants correctly perceived that the dialog informs them only of a “potential” threat, a couple participants misinterpreted this notification as informing them of a confirmed threat, as with one participant who believed that “*someone was hacking my account*”.

Qualitative analysis of participant responses regarding their decision to perform (or not perform) the privacy check showed participants weighed risks with response costs and made reasoned choices. These results roughly match those of the Mechanical Turk participants who evaluated our candidate designs. Participants who felt performing the privacy check was important reported that this stemmed out of a desire to confirm the validity of the reported risk or because they believed it better to be safe than sorry. Those who did not feel the privacy check worth doing, on the other hand, had either deemed the risk minimal or decided that conducting the privacy check was too inconvenient.

5.1.4 Privacy check

As with the authentication ceremony in the first user study, participants in our final user study may have seen and interacted with the privacy check screen without having conducted the privacy check itself. 17 of our 30 participants reported having seen the privacy check screen, with 3 participants unsure. 6 participant pairs fully performed the privacy check, while 3 participant pairs partially performed the check (one participant in each pair incorrectly informed their partner that they had already matched the identifiers and that they thus did not need to complete the full process). This is in contrast with the 5 pairs (out of 45) who performed the authentication ceremony in our first study.

These three “successful” misunderstandings had the same root cause—our design was not robust against false positives. Our design pops up an informative success dialog when a user

Table 2: Comparison of participant understanding of the authentication ceremony using Signal and our redesign.

Auth. cerem.	Signal		Redesign	
Correct	Verifies security (conversation)	2	Verifies security (conversation)	7
	Verifies device	2	Verifies device	1
		4 [16%]		8 [50%]
Partially correct	Verifies person (not impersonator)	8	Verifies security (connection)	3
	Verifies security (connection)	1	Verifies person (not impersonator)	2
	Prevents interception (connection)	1	Prevents interception (conversation)	2
	Improved security (conversation)	1	Verifies security	1
		11 [44%]		8 [50%]
Incorrect	Don't know	5		
	Verifies phone number	2		
	Verifies security (phone)	1		
	Prevents robocalls	1		
	Makes the contact trusted	1		
		10 [40%]		0 [0%]
Matching	Signal		Redesign	
Correct	Interception not possible	1	Interception not possible	6
	Verifies device	2	Verifies device	2
	Verifies security (conversation)	4	Verifies security (conversation)	1
		7 [26.9%]		9 [56.3%]
Partially correct	Verifies person (not impersonator)	7	Verifies person (not impersonator)	3
	Improved security	3		
	Prevents interception	2		
	Prevents interception (conversation)	2		
	Prevents interception (connection)	1		
	Verifies security (connection)	2		
		17 [65.4%]		3 [18.8%]
Incorrect	Don't know	1	Don't know	3
	Confusion	1	Confusion	1
		2 [7.7%]		4 [25%]
Non-matching	Signal		Redesign	
Correct	Interception occurring (MITM)	1	Interception occurring (MITM)	4
	Interception occurring	4	Interception occurring	2
	Conversation not secure	2	Conversation not secure	2
			Device is impersonator	1
		7 [28%]		9 [56.3%]
Partially correct	Interception occurring (contact is impersonator)	8	Interception occurring (connection not secure)	1
	Connection not secure	3	Interception occurring (contact is impersonator)	1
			Connection not secure	2
		11 [44%]		4 [25%]
Incorrect	Don't know	3	Don't know	2
	App is not secure	2	Conversation is secure	1
	Robocalls	1		
	Technical issues	1		
		7 [28%]		3 [18.8%]

taps the “Match” button. Unfortunately, this confused these participants who had mistakenly tapped the “Match” button. More specifically, one participant assumed that the “Match” button would activate an automated mechanism that would perform the verification for them. When the success dialog popped up in response, this participant assumed that the result had been in response to this “automated process”. The other mistaken participants accidentally tapped the “Match” button and were similarly misled by the resulting success dialog.

Table 2 shows a qualitative analysis of participant responses when asked the purpose of the privacy check, and the meaning of a matching or non-matching result, with responses coded and then categorized as correct, partially correct, and incorrect. This table reveals that comprehension of the purpose of the authentication ceremony and of the significance of matching and non-matching numbers visibly improved with our redesign. While far from perfect, these results are promising given the context: a non-sensitive task scenario, no accompanying instruction or tutorials, and no incentive. Risk communication was limited to the messages contained within the application.

For all categories and for both user studies, partially correct responses center on the same few misconceptions: believing the verification process itself to be an active prevention mechanism, believing the “connection” and not the conversation to be the entity to be secured, and believing that the verification process verified the contact’s identity, and not their device.

6 Discussion

6.1 Risk communication gives users the ability to make personal trade-offs between perceived risk and response cost.

Simply knowing that a negative outcome is likely to happen is not a sufficient reason to take action to prevent it: it must also be negative *enough*. As the participant quoted in the title of this work so eloquently stated, sometimes “*something isn’t secure, but I’m not sure how that translates into a problem.*” Indeed, this view was shared by a number of participants of our studies. We observed numerous instances where participants did not believe that conducting the authentication ceremony was a worthy use of their time, whether because they perceived their communications as non-sensitive and thus unworthy of protecting, or because they felt that performing the authentication ceremony would be too inconvenient. One shared response captures both these sentiments perfectly, “*If it was easy enough I would be happy to secure my conversation, but at the same time, how necessary is it?*”

While lowering response cost seems a natural way forward, particularly with automation, the deeply personal way in which calculations of risk function are made suggests obstacles ahead. Perceptions of risk severity in common scenarios will differ from person to person as a function of personal priorities and values. To wit, while many participants viewed communicating about our toy scenario as inherently non-sensitive, some participants were nevertheless uncomfortable at the realization that interception was “occurring”. One such participant, commenting on the thought of an interceptor eavesdropping on their discussion of a fictitious Hawaii trip, remarked, “*Even though it’s only about fish, that’s not really cool with me.*” We thus observe differences in risk assessment from different individuals although both the type of informa-

tion being communicated and the nature of the threat itself were identical in all cases.

For these reasons, it is our position that enabling users to truly make informed decisions requires properly communicating the nature and likelihood of the risk and the cost of recommended protective measures, and then giving them the freedom to determine that *not* actively protecting themselves is actually the decision most in line with their interests.

6.2 Users' strategies for coping with online threats extend beyond the ecosystem of your app.

Although our redesign evidenced both higher rates of participants conducting the authentication ceremony as well as comprehension, participants' responses of how they might mitigate the risks they had perceived did not change in any notable fashion. Despite both having been made aware of a protective measure (in the privacy check) and also having understood its purpose, participants ultimately did not find it a reliable measure for mitigating a perceived interception risk should they encounter a similar situation in the wild. Rather, participants mentioned self-filtering, restricting app permissions, and using alternative apps or channels of communication as key strategies for dealing with the interception threat introduced in the study.

This appears to be due to varying ideas about the source of the risk; in-app mitigating measures can only be depended on to do so much. Because the privacy check and associated messaging only informed users that conversation confidentiality had been violated, but not *how* that interception had been accomplished, users completed the process of threat assessment with personal interpretations of the origin of the interception risk. Relevantly, if the source of the risk is perceived to be outside the scope of the app—or even the app itself—it seems imprudent to rely on mitigating strategies that fall within the domain of the app.

System trust, perhaps unsurprisingly, appears to play a key role in this calculation. One participant response as to how they might better protect themselves is particularly ironic—they would forego use of Signal and “*use [a] secure messaging app like Facebook Messenger*”. Facebook Messenger does not protect conversations with end-to-end encryption by default, unlike Signal. However, due to unfamiliarity with Signal, and trust in Facebook, this participant's preferred strategy would be to move from a secure messaging platform to a less secure one.

Future work could examine whether additional risk communication regarding the source of the threat could lead to improved understanding of the efficacy of the privacy check. System designers should also consider that users choose, to varying extents, appropriate responses to perceived threats, and that these include viable methods above and beyond what the system itself offers.

7 Limitations

Our cognitive walkthrough was thorough but limited to the expertise of the authors who participated in it. We ameliorate this by having a variety of backgrounds among those who participated, but a walkthrough performed by other experts or novices may find different issues with Signal's notifications. Our Mechanical Turk studies of icons and notification flows are limited to a simulated experience and thus may not match what users would feel or choose when interacting directly with the application. Our lab studies were limited to a young, college student population and may not generalize to a larger or more diverse population. Our Mechanical Turk results from the simulation provide some evidence that the results of the second lab study generalize to a larger, more diverse population. It would be helpful to study populations with different risk-cost trade-offs, such as immigrants or dissidents, and to ascertain that risk communication translates well to other cultures and languages. Our lab studies are also limited because users may act differently due to the Hawthorne effect [24]. Several participants made comments indicating this limitation was present, such as “*while it is very concerning to me that someone could be intercepting my conversation, I thought that it was just because it was in a study.*” However, because the focus of our study was on comprehension as opposed to behavior, this effect may be less impactful in our study.

Aside from these more common issues, we also observed a bug in Signal's phone call functionality. The first time a Signal user makes an outgoing phone call, the caller is unable to hear audio although the recipient can hear clearly. Participants in our study simply redialed their partner when this occurred, typically chalking the issue up to a spotty wireless connection. This error, however, was present in the user studies evaluating both the original version of Signal and our redesign, so if this bug did have an effect, it likely existed in both cases, and thus is unlikely to have caused discrepancies in our results.

8 Conclusion

In this paper, we present the results of our experience redesigning the risk communication surrounding Signal's authentication ceremony for comprehension. Our three-part process reveals significant obstacles to understanding in Signal's current design, and demonstrates the effectiveness of applying risk communication principles to system design. Our user studies, which deliberately employ a non-sensitive communication task, provide evidence that users' decisions *not* to enact protective behaviors are actually conscious, informed decisions that are the product of balancing response cost and risk assessment. We further find that users rely on a host of protective behaviors that exist beyond the scope of any particular app or system, and that, consequently, responses to perceived threats may similarly exist outside of system designers' control.

9 Acknowledgments

The authors thank the reviewers and our shepherd for their helpful feedback. This material is based upon work supported by the National Science Foundation under Grants No. CNS-1528022 and CNS-1816929, and by the Department of Homeland Security (DHS) Science and Technology Directorate, Cyber Security Division (DHS S&T/CSD) via contract number HHSP233201600046C.

References

- [1] Ruba Abu-Salma, Kat Krol, Simon Parkin, Victoria Koh, Kevin Kwan, Jazib Mahboob, Zahra Traboulsi, and M Angela Sasse. The security blanket of the chat world: An analytic evaluation and a user study of Telegram. In *European Workshop on Usable Security (EuroUSEC 2017)*. Internet Society, 2017.
- [2] Piotr D Adamczyk and Brian P Bailey. If not now, when?: the effects of interruption at different moments within task execution. In *SIGCHI Conference on Human Factors in Computing Systems (CHI 2004)*. ACM, 2004.
- [3] Brian P Bailey and Joseph A Konstan. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior*, 22(4):685–708, 2006.
- [4] Robert Biddle, Paul C Van Oorschot, Andrew S Patrick, Jennifer Sobey, and Tara Whalen. Browser interfaces and extended validation SSL certificates: an empirical study. In *ACM Cloud Computing Security Workshop (CCSW 2009)*. ACM, 2009.
- [5] National Research Council et al. *Improving risk communication*. National Academies, 1989.
- [6] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. You’ve been warned: an empirical study of the effectiveness of web browser phishing warnings. In *SIGCHI Conference on Human Factors in Computing Systems (CHI 2008)*. ACM, 2008.
- [7] Michael Fagan and Mohammad Maifi Hasan Khan. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In *Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX, 2016.
- [8] Adrienne Porter Felt, Alex Ainslie, Robert W Reeder, Sunny Consolvo, Somas Thyagaraja, Alan Bettes, Helen Harris, and Jeff Grimes. Improving ssl warnings: Comprehension and adherence. In *SIGCHI Conference on Human Factors in Computing Systems (CHI 2015)*. ACM, 2015.
- [9] Vaibhav Garg and Jean Camp. Heuristics and biases: implications for security design. *IEEE Technology and Society Magazine*, 32(1):73–79, 2013.
- [10] Cormac Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *New Security Paradigms Workshop (NSPW 2009)*. ACM, 2009.
- [11] Hsiu-Fang Hsieh and Sarah E Shannon. Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9):1277–1288, 2005.
- [12] Clinton M Jenkin. Risk perception and terrorism: Applying the psychometric paradigm. *Homeland Security Affairs*, 2(2), 2006.
- [13] Doohwang Lee, Robert Larose, and Nora Rifon. Keeping our network safe: a model of online protection behaviour. *Behaviour & Information Technology*, 27(5):445–454, 2008.
- [14] James E Maddux and Ronald W Rogers. Protection motivation and self-efficacy: A revised theory of fear appeals and attitude change. *Journal of Experimental Social Psychology*, 19(5):469–479, 1983.
- [15] Daniel C McFarlane and Kara A Latorella. The scope and importance of human interruption in human-computer interaction design. *Human-Computer Interaction*, 17(1):1–61, 2002.
- [16] Jason RC Nurse, Sadie Creese, Michael Goldsmith, and Koen Lamberts. Trustworthy and effective communication of cybersecurity risks: A review. In *Workshop on Socio-Technical Aspects in Security and Trust (STAST 2011)*. IEEE, 2011.
- [17] Fahimeh Raja, Kirstie Hawkey, Steven Hsu, Kai-Le Clement Wang, and Konstantin Beznosov. A brick wall, a locked door, and a bandit: a physical security metaphor for firewall warnings. In *Symposium on Usable Privacy and Security (SOUPS 2011)*. ACM, 2011.
- [18] Rob Reeder, E Cram Kowalczyk, and Adam Shostack. Helping engineers design NEAT security warnings. In *Symposium On Usable Privacy and Security (SOUPS 2011)*. ACM, 2011.
- [19] Ronald W Rogers. A protection motivation theory of fear appeals and attitude change. *The Journal of Psychology*, 91(1):93–114, 1975.
- [20] Angela Sasse. Scaring and bullying people into security won’t work. *IEEE Security & Privacy*, 13(3):80–83, 2015.

- [21] Svenja Schröder, Markus Huber, David Wind, and Christoph Rottermann. When SIGNAL hits the fan: On the usability and security of state-of-the-art secure mobile messaging. In *European Workshop on Usable Security (EuroUSEC 2016)*. IEEE, 2016.
- [22] Paschal Sheeran. Intention—behavior relations: a conceptual and empirical review. *European Review of Social Psychology*, 12(1):1–36, 2002.
- [23] Paschal Sheeran and Thomas L Webb. The intention—behavior gap. *Social and Personality Psychology Compass*, 10(9):503–518, 2016.
- [24] Andreas Sotirakopoulos, Kirstie Hawkey, and Konstantin Beznosov. On the challenges in usable security lab studies: lessons learned from replicating a study on SSL warnings. In *Symposium on Usable Privacy and Security (SOUPS 2011)*. ACM, 2011.
- [25] Cheri Speier, Joseph S Valacich, and Iris Vessey. The influence of task interruption on individual decision making: An information overload perspective. *Decision Sciences*, 30(2):337–360, 1999.
- [26] Joshua Sunshine, Serge Egelman, Hazim Almuhiemedi, Neha Atri, and Lorrie Faith Cranor. Crying wolf: An empirical study of SSL warning effectiveness. In *USENIX Security Symposium*. USENIX, 2009.
- [27] René van Bavel, Nuria Rodríguez-Priego, José Vila, and Pam Briggs. Using protection motivation theory in the design of nudges to improve online security behavior. *International Journal of Human-Computer Studies*, 123:29–39, 2019.
- [28] Elham Vaziripour, Justin Wu, Mark O’Neill, Ray Clinton, Jordan Whitehead, Scott Heidbrink, Kent Seamons, and Daniel Zappala. Is that you, Alice? a usability study of the authentication ceremony of secure messaging applications. In *Symposium on Usable Privacy and Security (SOUPS 2017)*. USENIX, 2017.
- [29] Elham Vaziripour, Justin Wu, Mark O’Neill, Daniel Metro, Josh Cockrell, Timothy Moffett, Jordan Whitehead, Nick Bonner, Kent Seamons, and Daniel Zappala. Action needed! helping users find and complete the authentication ceremony in Signal. In *Symposium on Usable Privacy and Security (SOUPS 2018)*. USENIX, 2018.
- [30] Rick Wash, Emilee Rader, and Chris Fennell. Can people self-report security accurately?: Agreement between self-report and behavioral measures. In *SIGCHI Conference on Human Factors in Computing Systems (CHI 2017)*. ACM, 2017.
- [31] Irene Woon, Gek-Woo Tan, and R Low. A protection motivation theory approach to home wireless security. *International Conference on Information Systems (ICIS 2005)*, 2005.
- [32] Michael Workman, William H Bommer, and Detmar Straub. Security lapses and the omission of information security measures: A threat control model and empirical test. *Computers in Human Behavior*, 24(6):2799–2816, 2008.

A Signal authentication flow

Figure 4 shows a flow diagram of different screens in Signal when the encryption key changes for a contact, along with the transitions between screens based on user input. The top path is the “message not delivered” flow, which appears to send a message but shows a status indicating that the send failed. The bottom path is the “message delivered” flow, which only shows a notification but otherwise proceeds normally. The middle path is the “message blocked flow”, which prevents the user from sending a message initially.

B Redesigned Signal authentication flow

Figure 5 shows the redesigned authentication flow. There is only a single path, using a blocked message dialog along with a shield message in the conversation log.

Figure 6 shows the new notifications that correspond to this flow. If the user attempts to send a message after the encryption keys have changed, the message is blocked and a privacy check dialog is shown (upper left). From here, if the user taps “Get Started”, they proceed to the privacy check screen (top middle). They can use either the phone call (top right) or QR code scanner (bottom right). They can choose “Not Now” from either the privacy check dialog or the privacy check screen, and they will proceed to the reminder dialog (bottom left). The result of the privacy check (failure or success) is shown in Figure 7.

Figure 8 shows the notifications in the conversation log. First, when encryption keys change, a notification is displayed that recommends a privacy check (Figure 8a). Later, if the user completes the privacy check, a different notification is shown if the identifiers match (Figure 8b) or don’t match (Figure 8c). These notifications scroll as new messages are added to the conversation.

C Privacy check indicators

Figure 9 shows the new privacy check indicators. Tapping on the indicator brings up the corresponding privacy check screen, depending on the current state of the conversation, as shown in Figure 10. These same screens are accessed if a user taps of any the conversation log notifications.

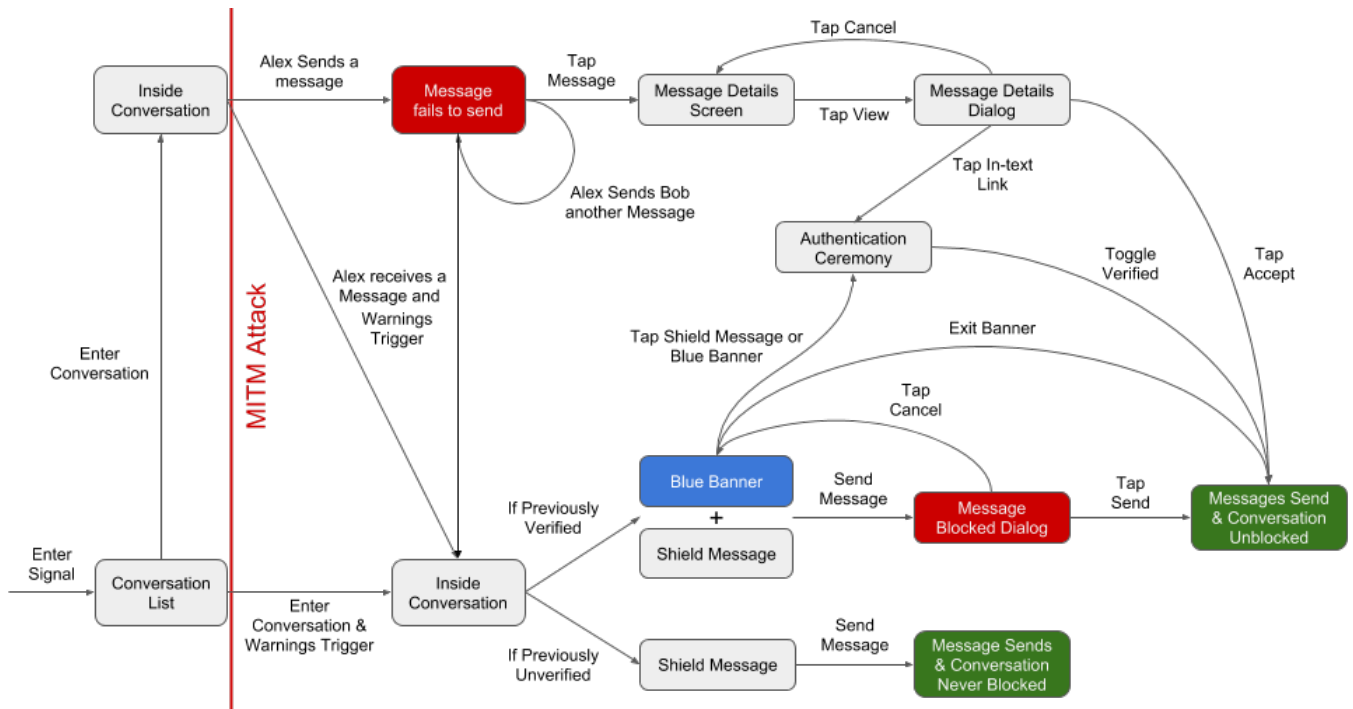


Figure 4: Flow diagram depicting the how Signal reacts to a safety number change.

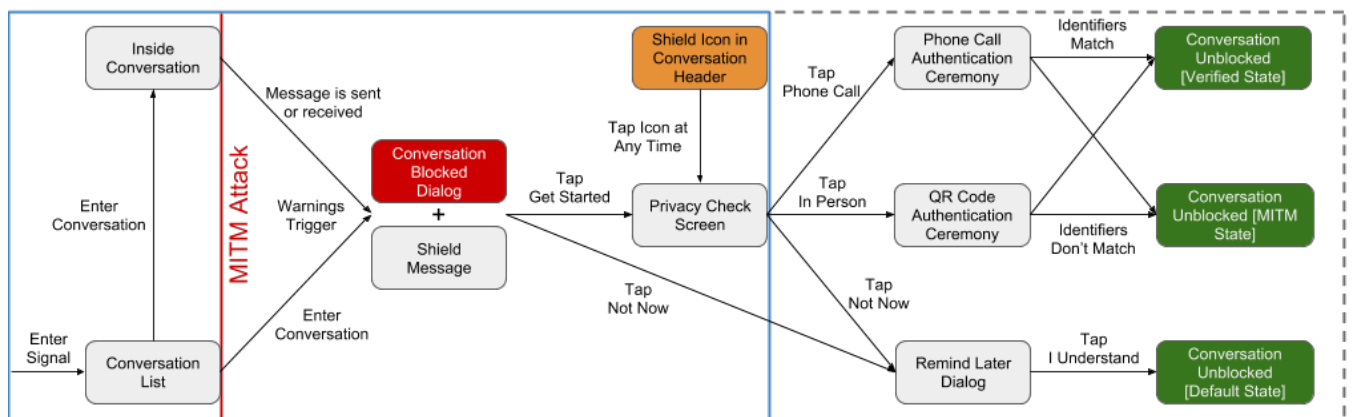


Figure 5: Flow diagram depicting how our redesigned Signal reacts to a safety number change. The blue box encloses the elements and choices with analogous equivalents in the original Signal client. The area contained by the dashed lines shows choices, elements, and state changes that we added in our version that are expansions on the authentication ceremony and beyond.



Figure 6: Privacy check notification flow

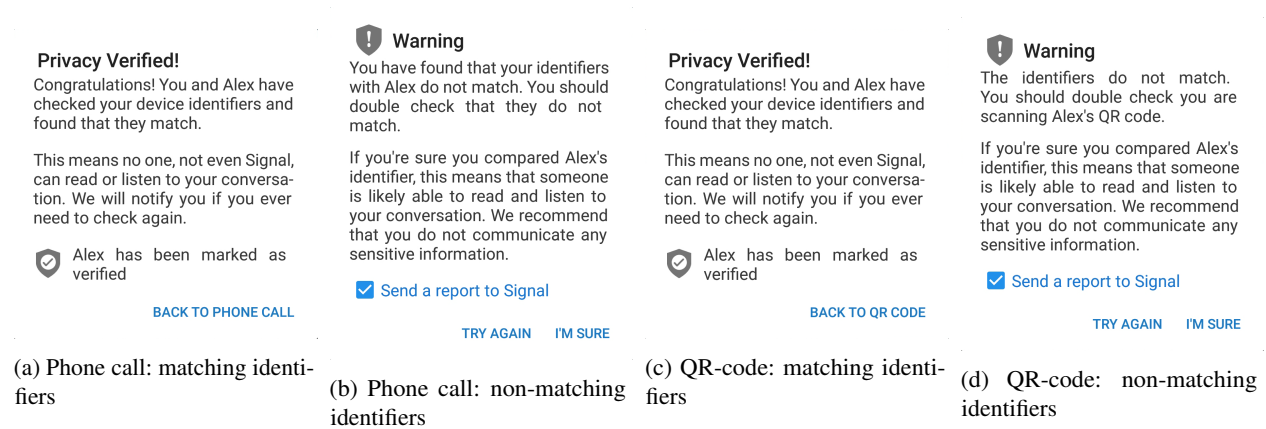


Figure 7: Phone call and QR code privacy check results

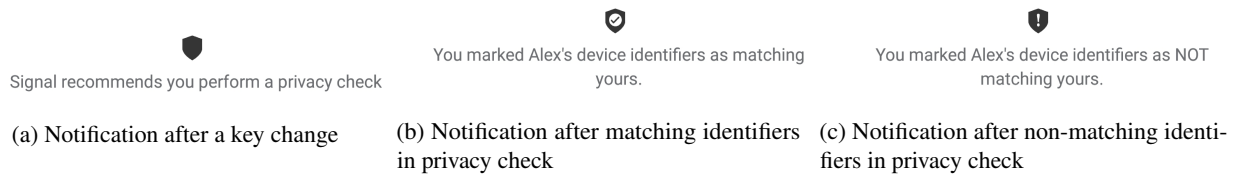


Figure 8: Conversation log notifications

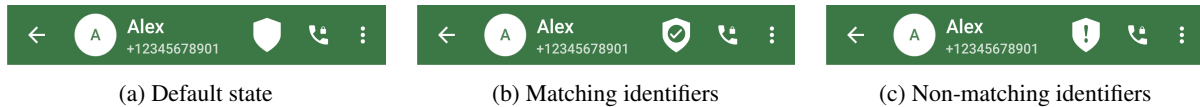


Figure 9: Privacy check indicator

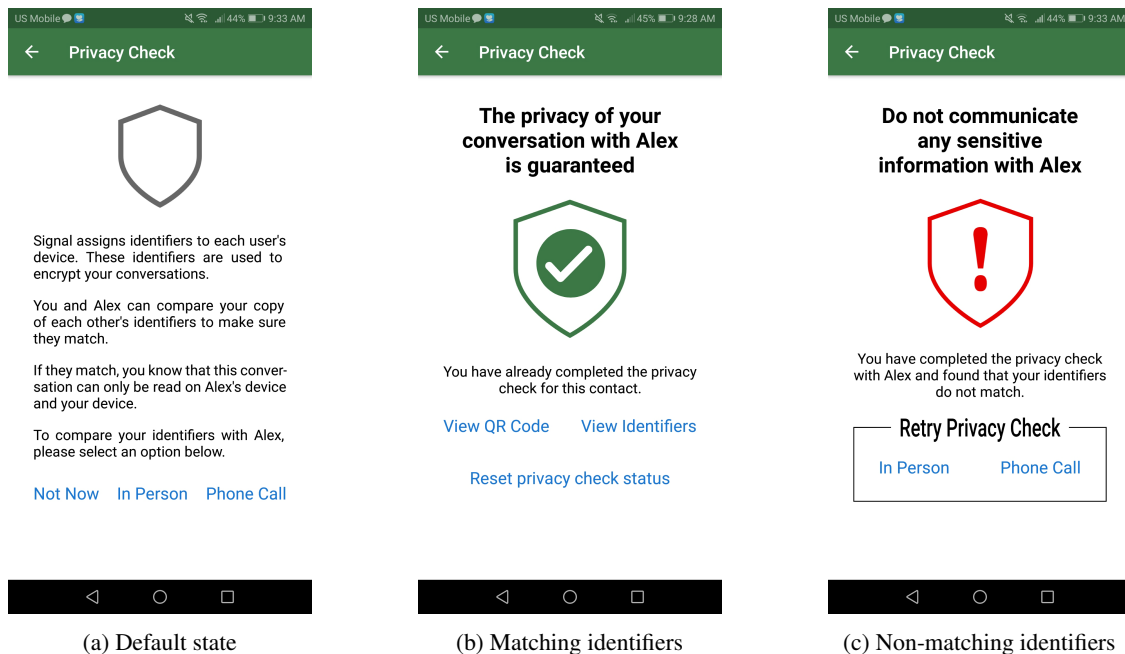


Figure 10: Privacy check screen