Can Johnny be a whistleblower? A qualitative user study of a social authentication Signal extension in an adversarial scenario

Maximilian Häring and Julia Angelika Grohs, *University of Bonn;*
Eva Tiefenau, *Fraunhofer FKIE;* Matthew Smith, *University of Bonn and Fraunhofer FKIE;*
Christian Tiefenau, *University of Bonn*

https://www.usenix.org/conference/soups2024/presentation/haring

This paper is included in the Proceedings of the Twentieth Symposium on Usable Privacy and Security.

August 12–13, 2024 • Philadelphia, PA, USA

978-1-939133-42-7

Open access to the Proceedings of the Twentieth Symposium on Usable Privacy and Security is sponsored by USENIX.

Can Johnny be a whistleblower? A qualitative user study of a social authentication Signal extension in an adversarial scenario

Maximilian Häring
University of Bonn

Julia Angelika Grohs
University of Bonn

Eva Tiefenau
Fraunhofer FKIE

Matthew Smith
University of Bonn, Fraunhofer FKIE

Christian Tiefenau
University of Bonn

Abstract

To achieve a higher level of protection against person-in-the-middle attacks when using common chat apps with end-to-end encryption, each chat partner can verify the other party's key material via an out-of-band channel. This procedure of verifying the key material is called an authentication ceremony (AC) and can consist of, e.g., comparing textual representations, scanning QR codes, or using third party social accounts. In the latter, a user can establish trust by proving that they have access to a particular social media account. A study has shown that such social authentication's usability can be very good; however, the study focused exclusively on secure cases, i.e., the authentication ceremonies were never attacked. To evaluate whether social authentication remains usable and secure when attacked, we implemented an interface for a recently published social authentication protocol called SOAP. We developed a study design to compare authentication ceremonies, conducted a qualitative user study with an attack scenario, and compared social authentication to textual and QR code authentication ceremonies. The participants took on the role of whistleblowers and were tasked with verifying the identities of journalists. In a pilot study, three out of nine participants were caught by the government due to SOAP, but with an improved interface, this number was reduced to one out of 18 participants. Our results indicate that social authentication can lead to more secure behavior compared to more traditional authentication ceremonies and that the scenario motivated participants to reason about their decisions.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2024.
August 11–13, 2024, Philadelphia, PA, United States.

1 Introduction

End-to-end encryption (E2EE) is a well-known and broadly applied technology in messaging apps. Its implementation helps to improve the privacy of billions of people. However, E2EE cannot provide authenticity without the interaction of users. To have authenticity, chat partners must ensure that the correct key material is used, i.e., the service provider is not tampering with the keys to mount a person-in-the-middle (PITM) attack.

The task of comparing the key material of the communication partners, e.g., by meeting in person and showing them, is called an *authentication ceremony (AC)*. By correctly carrying out an AC, users can be sure that they are talking confidentially with the right person. However, the default in current messaging apps is to trust the first keys given to users by the provider without encouraging an AC [2] and inform users when these keys change. Studies show that few users run authentication ceremonies, and many users do not know the cryptographic notion of authentication and how to handle the corresponding ceremonies [3, 5].

A possible reason why few users have a reason to verify keys is that even without verification E2EE provides a good level of protection as mass surveillance is resource-hungry and disincentivized for the attacker; getting caught is fairly likely due to key-change notifications that can be noticed by the provider or experts, e.g., facilitated by key transparency [9, 13, 28]. However, targeted surveillance can still be a threat as it is technologically possible, and the risk-benefit ratio for the attacker could be worthwhile. Consequently, we believe that if there is a need for authentication ceremonies, it is most pressing in high-risk scenarios, e.g., when one is a political dissident, a whistleblower, or a government employee. While the single tasks that are necessary for ACs can be done quickly and with rather low false-acceptance rates [18, 25], studies provide evidence that current authentication ceremonies are difficult and error prone [5, 6, 17, 27].

A fairly new solution for remote[1] authentication, "social

[1] "Remote" refers to a setting where the two communication partners carry

authentication (SA)" was suggested by prior research and leverages social networking sites as a trust anchor [8, 11, 24].

The idea behind this solution is to reduce the verification task to something users can already do and intuitively grasp. For SA, users do not need to compare key material directly; instead, they must decide which identity provider, e.g., a social media site, to trust and recognize an already known account. As such, users need to have knowledge about the contact they want to authenticate and know their identifier (e.g., Alice42) on the chosen identity provider (e.g., facebook.com). Vaziripour et al. [24] tested the concept in a laboratory study and found the approach to have good usability. They reported that participants found the concept convenient and matched "how participants thought of verification." However, their solution was tested under ideal conditions, i.e., without any attackers. Nevertheless, the researchers noted that SA makes identity spoofing and impersonation attacks possible. Currently, no work on SA in an attack scenario exists. To fill this knowledge gap, we conducted a user lab study where we simulated an attack scenario and compared SA to the already established ACs of key fingerprint comparison and QR codes.

This work contributes a novel methodology for comparing ACs and an interface to make social authentication similarly usable as safety numbers or QR codes. We extend the existing literature on ACs and how they are researched by testing an attack scenario in a **user study** of a SA approach. We were especially interested in the participants' reactions toward impersonation attacks, i.e., how often they would notice the attack and how they would proceed with a given task.

We created a scenario that resembles, more closely than previous work, a realistic use case for users needing an authentication method. To motivate the participants to authenticate and mimic real-world situations, they had to act as whistleblowers in an authoritarian regime and contact journalists. This **study design with a scenario with reasonable participant motivation** allowed us to observe the entire process of the authentication ceremonies. In contrast to Vaziripour et al.'s study [24], which proposed a form of SA, Linker et al. [11] formally defined SA and presented a protocol with proven security properties. They called the protocol SOAP and developed a prototype that worked, with limitations, within the current internet eco-system. This means our results can be directly applied to their prototype and hopefully increase the security of users.

During our analysis, we were guided by the following research questions:

RQ1 - Detection: How resistant is SOAP to impersonation attacks?

RQ2 - Reaction: How do participants react to a detected impersonation attack?

RQ3 - Perception: What are users' perceptions of SOAP (usability, trustworthiness), with a focus on identity providers?

out an AC without meeting in person. Although we phrase ACs as a task for two users, it often works similarly for more than two.

In a pilot study, nine participants used a simple interface based on the protocol proposed by Linker et al. [11], which was implemented as an extension to the Signal app [19]. Many participants failed to use SA correctly when under attack. After analyzing the results, we improved the interface and recruited 18 participants. Although our design improved the results so that only one participant behaved insecurely because of SOAP, six of the lab study's 18 participants failed to detect a PITM for other reasons. If applied to the real world, this would mean that they would be in danger if they were to rely on a tool like Signal for confidentiality.

The rest of this paper is structured as follows: In Section 2, we provide a short overview of relevant authentication methods, their shortcomings, and the concept of SA. In Section 3, we present the user study, and in Section 4, we discuss implications and further directions for research and messaging app developers.

2 Related Work and Background

In this section, we summarize ACs in the messaging app domain and related work about them to put social authentication into context.

2.1 Authentication Ceremonies

Comparing key material, a process called authentication ceremony (AC), has scarcely changed in the last few years. Via an AC, a PITM attack can be detected, e.g., if the attacker uses a key substitution attack [6].

Material for comparison is always based on the public key, but the visualization differs among apps [2, 6]: Signal initially displayed two public key fingerprints before changing to concatenation and currently displays a single safety number [14]. In addition to that, Signal also offers a QR code, which is a different representation of the single safety number. A recent version of Telegram (iOS 10.3.1) shows a scannable identicon, similar to a QR code, and a hex notation "generated from hashes of the DH secret chat keys" [22]. During phone calls, emojis are shown [21].

The success of an AC has its challenges. Herzberg et al. [6] structured these as deciding that a ceremony is needed, finding the ceremony in the user interface, executing the ceremony, understanding the result, and acting on it.

As it is assumed and evidenced [17, 26] that users struggle with ACs, studies looked at each of the steps in the process and tried to improve them. Vaziripour et al. [25] worked on guiding users to the ceremony interface. With opinionated design, they were able to lead 90% of their study participants to the ceremony. Wu et al. [27] worked on users' comprehension of safety number change notifications and found a need to communicate the possible risk to users as a motivation and basis to decide. Shirvanian et al. [18] studied whether the comparison act itself could be a problem. They found

evidence that in a remote setting (i.e., when users do not sit next to each other), comparison can be an error-prone task, mainly because users need to compare codes between two apps on the same device, with the need to remember the code. Tan et al. [20] and Livsey et al. [12] researched how different visualizations impact the comparison act. Although Livsey et al. found that their participants did not make many mistakes, Tan et al. found that visualization can greatly impact the outcome in an attack scenario, with success rates for the attacker varying between 6% and 72%. All these studies and the methods used rely on the same AC principle: a direct manual exchange and comparison of key material to authenticate the communication partner. As described in the next section, social authentication relies on a different principle.

2.1.1 Social Authentication

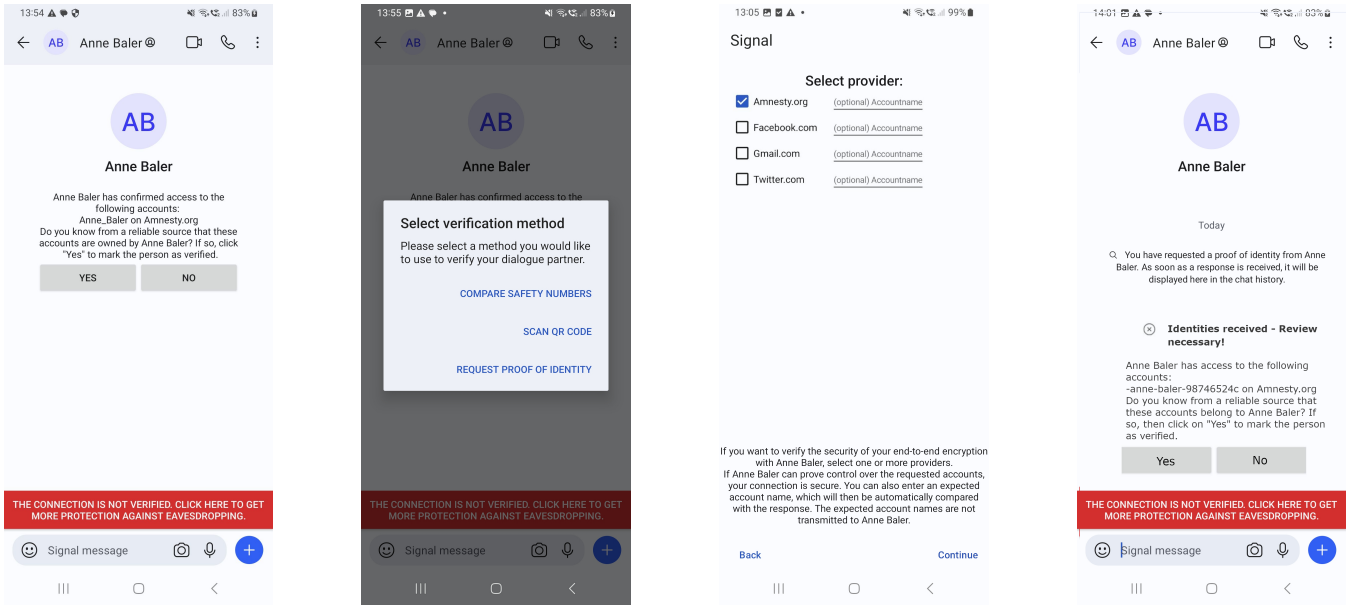
In the literature, two different topics are referred to as social authentication. According to Jain et al. [7], SA describes when Alice wants to log in to a service and another user, Bob, who is connected to Alice on the (social media) platform, is asked whether they are allowed to. This can be triggered, e.g., as a step in a risk-based authentication scheme. However, Vaziripour et al. [24] described SA as an AC completed through "social media." In SA, public key material is distributed through a social media provider. In this paper, we refer to this second notion of SA as an AC.

Concept With this AC, the challenge of the ceremony is shifted from selecting a secure channel, exchanging the key material, and comparing the fingerprints to deciding what provider to trust and recognizing an identifier.

An early application where this notion of SA is in place is Keybase. On Keybase, a user can provide proof of having access to an account by posting material on it. Afterward, other Keybase users can decide whether proof of access to that account is enough for them to identify the person [8]. Vaziripour et al.'s [24] proposed system is very similar. The researchers envisioned that Signal users would log in to their social media accounts during configuration, and the public key material would be posted there. Similar to the scheme utilized by Keybase, this would allow observers to see the material. For example, if Alice wants to check whether the E2EE on Signal is PITM-free and authentic, they could check whether Bob has provided a reference account on a trusted social media platform, in the following called identity provider (IdP). As the key material is posted online, it can be compared automatically and asynchronously. The decision Alice has to make is whether they trust the IdP and whether the account provided by Bob belongs to the person they want to contact.

Studies Vaziripour et al. [24] tested their idea in a lab study (21 participant pairs) and an online survey (N=421). They let the participants communicate via Signal and, if needed, guided them to the AC. Here, the participants were able to choose between three verification methods: *social media* (social authentication), *in person*, and *phone call*. The participants were allowed to choose from all three methods and were asked to use the remaining two after selecting one. The researchers found that the *social media* verification method had the best Single-Ease-Question (SEQ) score but was less trusted than the *in person* and *phone call* methods. Additionally, the participants chose the *in person* method first (n=20) more often than the *social media* method (n=12). The average configuration time of the *social media* method was 2 minutes and 32 seconds. On average, verification (which, in this case, meant looking at profile names and pictures) took 34 seconds. Varziripour et al. concluded that social media was not perceived as a highly trustworthy provider of authentication, but the participants liked the asynchronicity, that it worked remotely, and that it was partially automated. As the challenge of the AC changes, so does the attack surface. The participants in Varziripour et al.'s study mentioned the attack vector of fake profiles, which indeed seems to be a major challenge for SA. Additionally, the key material has to be public. This could be problematic for some users due to privacy considerations.

SOAP Another recent proposal, "SOAP" [11], mitigates the need to have the key material public and aims to find a way to bootstrap SA in the current internet without too much effort from the provider's site. Hence, SOAP utilizes IdPs, not necessarily social media providers. An identity provider (IdP) could be any entity providing an OpenID Connect service, hoping for a relatively fast and easy adoption. If Alice wants to check the security of the chat with Bob, Alice asks Bob to prove that they have control over an account at a specific (listed) IdP. This can be done by just sending a message to Bob to do so or utilizing a UI flow as proposed in this paper. After that, Bob's client asks an IdP to sign the chat's safety number in combination with his account on the platform. Bob has to first log in to the provider before the IdP signs a request. The signing is done automatically. Bob then forwards the signed message returned from the IdP to Alice. Afterward, Alice has a statement from the IdP that says: With whom you are talking to, identified by this safety number, has control over account "XYZ" on my platform. More technically, Bob's client starts an OpenID Connect authorization code flow with the salted hash of the safety number and a nonce. The resulting ID token (including the nonce) and the salt are forwarded to Alice. Alice's client can now check whether the safety number it has matches the one incorporated within the token. Alice then has to decide whether the identity provided is as expected and wanted. To the best of our knowledge, currently, no research on SA attack scenarios exists. We fill this gap in the remainder of this paper.



(a) New chat for participants in the **pre-registration** condition. They saw a non-requested SOAP answer. Otherwise, the chat was empty. The **red banner** nudged the participant to find the ceremonies.

(b) Menu that opens if participant clicks the red banner. The first two options led to the currently implemented safety number site with slightly modified text. The third option opens a SOAP request interface.

(c) A SOAP request can be made by selecting an IdP and, if wanted, adding an expected identifier. By continuing, a SOAP request is sent. If an identifier is added, the recipient cannot see this.

(d) If it cannot be determined automatically whether the identifier is correct, the user must decide.

Figure 1: Translated screenshots of the UI used for the lab study.

3 User Study

We tested SOAP in a lab study with a preceding pilot study. We implemented and started with a simple SOAP [11] interface for the Android Signal app. Based on a pilot study (n = 9), we adapted the interface. At last, we ran a lab study with 18 participants. This section describes the resulting SOAP interface and the study design and presents results from the pilot and the lab study.

3.1 Technical Implementation - UI

Linker et al. [11] presented with the protocol an accompanying prototype that implements the technical protocol but does not hint at its capabilities to the user. Only one button in the app's share menu suggested the existence of SA. So, the verifier has to know that SOAP exists and somehow agree with the to-be-verified person what identities and IdPs are available and then ask for proof. For this study, we were not interested in whether people could find the icon, and we did not want to explain the idea in a workshop. Seeing that the prototype's design was not ready for our purposes, we adapted it to the needs of our study. We used the Signal app because the prototype builds on it, it is open source, and previous studies also used Signal. In the following sections, we detail relevant elements of the technical implementation from after the pilot study. An earlier version of the interface can be seen in the Appendix (Figure 3).

3.1.1 Hint to the Ceremonies

To test SA, we wanted to point the participants directly to the relevant parts of the interface. Vaziripour et al. [25] successfully led users to the ACs with a clear, visible red banner above the text entry field in chat views, and we adopted the same method (see Figure 1a). A click on the banner triggered a dialog with the three ceremonies (see Figure 1b): Safety Number, QR-Code, and SOAP.

3.1.2 QR Code and Safety Number

When the participants clicked on the QR code or safety number button, they landed on the slightly modified safety number page in the chat settings as in recent versions of Signal (see Figure 5d in the Appendix), where a QR code could be scanned, and the chat's safety number read. A message must have been exchanged with the contact for a chat to have a safety number. If the participants tried to access this page

without prior communication, a popup reminded them that a first exchange must happen. Please note that although Signal provides a unique safety number per chat, it is only a concatenation of two per-user numbers. So, just the half belonging to the contact has to be checked. We added an explanation of this to Signal's settings page.

3.1.3 SOAP - The Social Authentication Protocol

Choosing "request proof of identity" opened a window to start the flow for SOAP [11] that asked the user to select an IdP and what accounts the chat partner should prove access to (see Figure 1c). The user could choose as many of the given IdPs as they wanted and optionally fill in their communication partner's expected account names on these platforms. At the time of designing the study, the original prototype only supported Microsoft and Gitlab. To provide more providers, we omitted the technical procedure, and the journalists just responded with a predefined formatted string interpreted by Signal as a valid response on the participants' side. After receiving the response (Figure 1d), users could mark the user as verified, or, if an identifier was pre-filled, the client automatically marked the response as correct/incorrect.

The interface's UI can also be seen in Figures 4 and 5 in the Appendix. The source code is available at `https://osf.io/dsyfr/`.

Pre-Registration Mode The pre-registration mode is a variant of SOAP not proposed by Linker et al. but invented by us based on observations in the pilot study (see Section 3.2.6). Instead of asking the chat contact to prove access to a selected IdP and waiting for the response, the chat contact provided this proof in advance by logging in to the IdP once. This way, a new chat with this contact shows a SOAP response without any previous message exchange (see Figure 1a). This mode is similar to the proposal of Vaziripour et al. [24], where the participants liked that they did not have to communicate with the chat partner to check their identity. Technically, this would be possible in the same way the provider's server shares the public keys, or the material could be posted publicly as proposed by Vaziripour et al. [24].

3.2 Methodology

We conducted a lab study where we tested the detection rate of and the reactions to an impersonation attack on a new AC. The documents for the study can be found in the Appendices.

3.2.1 Setting and Scenario

When developing our scenario, we looked at previous studies on ACs. Herzberg at al. [5] reported that participants recognized to act differently depending on the situation, e.g., based on the importance of a contact, and Wu et al. [27] discussed participants' need to be able to assess the need for an AC. Previous studies observing human behavior and ACs used very simple scenarios [18] or settings where there was little explicit (intrinsic or extrinsic) motivation for the participants to behave securely [5, 17, 24–27]. We wanted our participants to be motivated to conduct the AC, so we provided a scenario that gave them a reason to do so: a whistleblower scenario. We hoped the participants would understand the importance of being cautious, as they know the consequences of deanonymization, e.g., losing their job and reputation, prison, or even death. To check the realism of our scenario, we searched news sites and found examples where Signal was proposed as a channel for communication [1, 10, 15, 16, 23, 29].

In some previous studies with ACs, participants were invited in pairs [17, 26], sometimes knowing each other [26]; hence, they would have been able to judge whether the contacted person was the correct individual based on voice, looks, and behavior or meeting in person. We reduced these mitigating strategies through the scenario so the participants could not know the person they interacted with and could not verify the person via human characteristics.

Taking all this into consideration, we ended up with the following scenario outline: The participant, named Alex, is a whistleblower. Their colleague Hannah sent them documents revealing a political scandal via Signal. Their conversation was verified in person before receiving a .zip file containing sensitive data. Hannah is only reachable via Signal. Alex's task is to contact three journalists and send them the documents after ensuring they are interested in the data and the communication is safe. Alex receives information about these journalists on business cards (see Appendix A.6 for details). As part of the introduction, the participants were told that the business cards came from a trusted source. Communication could only occur through Signal's text function; other channels were not allowed. Each journalist had one intended possibility to be verified, for which we printed the necessary information on each business card: **Amira via safety number**, **Michael via QR code**, and **Anne via SOAP**. This way, the participants were nudged to use every method at least once. However, the participants were unaware that the authoritarian government of the scenario was suspicious of Alex and all connection attempts were attacked with impersonation attacks. Technically, this could be implemented by hacking the Signal servers and mounting a PITM, but also by possessing the SIM card, e.g., by a SIM swapping attack. So, all the verification checks failed: the safety number shown in Signal differed from the number and the QR code on the business card, and for each SOAP request, the provider or identity did not match (see Table 1). The only correct behavior for participants was to abort all communication attempts, which was explicitly allowed in the task description. The within-subject design allowed us to compare the available ACs and generate more observations with the small sample. We believed that

participants might buy into the scenario, but we want to note that although we described and explored a high-risk situation, the concrete setting lacked realism. We simplified by defining that the business cards Alex has available are to be trusted without exploring how realistic that is. Also, in a real-world scenario, we assume that Alex would compare the available information with further researched ones, e.g., an email address or a well-known social media account that can be confirmed from different sites. Also, deciding not to use Signal but to work via other channels is possible. To implement all possibilities realistically is sadly out of scope for a lab study and needs further research. More studies are needed to establish a best practice for real at-risk users. For this study, the most important was that the participants accepted the scenario as realistic and plausible.

SOAP Attackers Capabilities The responses to the SOAP requests were randomly selected from three cases:

- Wrong provider - correct identifier
- Correct provider - wrong identifier
- Currently no access available

If multiple IdPs were asked for in a single request, the cases were picked without duplicates, so that with three asked IdPs, there were three different cases. The concrete available answers can be seen on Table 1.

Provider	Correct Identifier	Identifier available to the attacker
Amnesty.org	anne-baler-98746524b	anne-baler-13885412b
Facebook.com	Anne_Baler	AnneBaler
Gmail.com	n.a.	anne-baler@gmail.com
Twitter.com	@AnneBaler	@AneBaler
Amnesty.org	n.a.	a.patel@amnesty.com
Facebook.com	n.a.	Amira_Patel_86
Gmail.com	n.a.	patel_amira_86@gmail.com
Twitter.com	n.a.	@apatel
Amnesty.org	n.a.	m.kobel@amnesty.org
Facebook.com	n.a.	Michael_Kobel
Gmail.com	n.a.	michael_kobel@gmail.com
Twitter.com	n.a.	@michael_kobel

Table 1: This table displays the identifier the attacker sent and what the correct one would have been. Providers without correct identifiers are marked as "n.a.". For these, the participant could not determine the correct identifier. Participants received one of three responses: a) no identifier, simulating no current access to the account, b) an incorrect identifier for the requested provider, or c) a known identifier that is correct but for a provider different from the one requested.

SOAP does not submit the identifiers the requester expects. So, the attacker does not know whether the requester filled in identifiers. In the interface of the pilot study, if a provider was requested and the response did not contain the provider, it was marked as a missing provider. The attacker could send an accompanying message like "Sorry, I currently have no access to this account," hoping the requester would not mind. In the more opinionated later interface, any deviation from the request was marked as a failure. So, for the second half of the participants, we changed the attacker. The attacker would always send some form of identifier, hoping that the requester had not filled in an expected identifier. If no identifier was filled in during request, the participants had to decide whether the identity submitted was sufficient.

3.2.2 Recruitment and Participants

We conducted a small pilot study (n = 4), recruiting participants from our research group's contacts. After this, we recruited 13 participants from an undergraduate usable security and privacy lecture and confronted them with an early interface. For their participation, the participants received bonus points for the lecture exam and a bonus cash reward. They started with €5, and if they securely transmitted the sensitive data to a journalist, they received an additional €5 for each journalist. The participants were told they lose everything if they got caught, e.g., by sending the data to the wrong person. As all journalists suffered an impersonation attack, no bonus could be earned. To eliminate any motivation to collaborate with fellow students, we paid each participant €20 and asked them to keep the study details confidential.

For the lab study, we recruited participants via a behavioral economics lab mailing list where studies can be distributed. To recruit 21 participants, an invitation was sent out to 3000[2] randomly picked mailing list receivers over 18 years old. The lab had a strict no-deception rule, so we had to change our initial reimbursement scheme. To keep a risk/reward payment scheme for motivation, the participants received a base pay of €15 and had the chance to receive an additional €9 (€3 for the correct decision for each journalist). The entire bonus cash reward would be lost if they made one wrong decision. We provided this reward to motivate the participants to contact as many journalists as possible and try the different authentication methods while behaving securely: weighing the risk of not sending the data and receiving less money versus sending the data and risking losing everything except the base pay. We hoped this would lead them to act cautiously and align their interests with the scenario. We followed that scheme with one exception: P2 did not send any message, being cautious that even a single message could be a problem, and thus stayed safe. To gather more information, we asked them to do so. While they later made a mistake, we paid out the bonus in full since their first behavior was safe.

In the lab study, eleven out of 18 participants did not use Signal before. Also, most of the participants (15) never checked the safety numbers of their contacts in any app. The ages ranged from 21 to 46 with a median of 24. One participant did not give their age. The pilot took place in July 2023, and the lab study in October 2023.

3.2.3 Ethics

We received IRB clearance for all studies and adhered to the German data protection laws and the GDPR in the EU.

[2]We had no control over how many people were contacted.

All participants consented to their participation and the use of the data for research purposes before participating. The participants were informed that they could terminate their participation at any time without negative consequences and that, in such a case, all the respective data collected up to that point would be deleted. The participants of the pilot study received bonus points for the lecture exam, which could also be obtained in other ways.

3.2.4 Study Protocol

The study was conducted in three parts, as described in the following.

Part 1 - Intro The participants read and signed the consent form. Afterward, they received the material and were instructed to read the scenario text. Each participant was handed a pen, paper, and a smartphone with Android 13 and our modified version of Signal installed. Additionally, we handed them the three journalists' business cards (see Appendix A.6) in random order to counter ordering effects. The journalists each had an existing phone number to enable Signal communication. Further details on the business cards were fictive to avoid selection bias based on a newspaper's familiarity or reputation. The participants had to answer a quiz questionnaire on the phone before starting the scenario (see Appendix A.3). The quiz consisted of seven questions about the scenario. The participants could answer the questions as often as necessary to get all the answers correct.

Part 2 - Scenario We asked the participants to think aloud while working on the task, audio recorded the whole procedure, and screen recorded the smartphone. Their task was to choose journalists and try to contact them securely. The researcher giving the briefing was present in the room during these steps and ended the scenario after about 30 minutes to keep the whole study under one hour. The researcher had the option to extend the time a few more minutes if a participant was in the final stage of sending or verifying. A second researcher who was not present in the room manned the journalists' Signal accounts. They had a playbook (see Appendix A.7) that was expanded in new situations. If a participant asked for a communication method other than Signal, this was denied, as is the case in real-world scenarios.

Part 3 - Outro When a participant told the researcher they were done or the study time was up, they needed to complete a survey (see Appendix A.2). After this, there was a short interview followed by a debriefing (see Appendix A.5).

3.2.5 Analysis

We used qualitative and quantitative data to capture the results. As per our research questions, we were interested in:

1. Who tries to authenticate via a ceremony? We assumed this would be everyone as we added the red banner [25].
2. Which provider is chosen on SOAP? We assumed that most of the identifiers on the cards would be used.
3. Do participants detect the attack via SOAP? We assumed that most of them would.
4. How do the participants react? We assumed that the participants who detected a failed ceremony would abort contact.
5. How many participants fail the task? From the overall tone in related work, we assumed a few would.

A researcher who was present at all but one participant's sessions used notes, transcripts, screen recordings, and survey results to extract the steps participants took and where the participants failed. The researcher started by marking all positions in the recording relevant to the research questions, e.g., when a method was used and when and how a decision was made. A scenario is understood as failed if a participant sent a file to at least one journalist.

3.2.6 Results - Pilot study

In this section, we briefly describe the results of our pilot study. From the 13 participants (computer science students, abbreviated CS in the following), we excluded the data of three due to UI bugs and another participant who stated that they knew the study design beforehand. The data from the resulting nine participants was analyzed further. A table summarizing the results can be seen in the Appendix (Table 4).

The UI and the scenario text seemed to work, as all participants except CS-7 started every AC at least once.

Even though we intended for each journalist to be authenticated with exactly one method (safety number, QR code, or SOAP), all the business cards were provided with at least an email address. Following this and as we allowed to use custom providers, SOAP was not only used for Anne but for other journalists as well, with the work email being the most frequently used IdP (see Table 3 in the Appendix).

Overall, four of the nine participants forwarded the data to at least one journalist. All but one failure in the scenario can be traced back to SOAP. Specifically, we identified three reasons for failure.

Typosquatting: Three participants did not notice the typosquatting attack in SOAP or assumed it was acceptable, e.g., CS-3 recognized a provider mismatch but decided that an email identifier can only be verified as an account name if access to it is available. They all correctly saw that the safety number and QR code were invalid. We assume a more sophisticated attacker could have fooled more users.

"Marking" makes it secure: CS-7 contacted every journalist with a cover story. Afterward, they clicked on the safety number site, marked the journalists as verified, and sent the data. While that initially seemed rather strange to us, CS-7 explained in the interview and survey that they expected the chat to be verified and encrypted after this action.

Trust Chaining: Additionally, two of the participants verified one journalist and asked this journalist for the safety number of another journalist. The attacker provided the number seen by participants shown by the client. With this, the participants even accepted journalists who were previously perceived as suspicious.

Changes Based on the Pilot Study Based on the results, we made several modifications to our study design and how users interacted with the SOAP interface:

1. We adapted the SOAP interface to reduce possible attack surfaces and automated what could be automated.
2. We added a link to start a SOAP request on the safety number site in the settings.
3. The red banner no longer disappeared after verifying the person but turned green. This allowed a more direct way to the setting page and clearly indicated the chat's status.
4. To reduce the attack surface for typo attacks and match the current technical landscape, we removed the option to ask for custom IdPs and reduced the number of providers.
5. Based on the participants' comments and Vaziripour et al. [24], we assumed that a non-social media company would be favored and seen as more trustworthy. Therefore, we added Amnesty.org as a provider option.
6. We reworked the visuals of the UI, fixed glitches, added more text, and added guidance to the interaction of the SOAP responses depending on the outcome, e.g., obstacles to send in the case of an incorrect response.
7. As some of the participants in the pilot study were afraid of sending even a single message, they did not trigger the AC. To test SOAP without user interaction, we added the pre-registration mode as a between-subject condition (see Section 3.1.3), to which half the participants were assigned (see Table 2). We halved this group again by the provider/identity pair they would see: half saw the identifier for Facebook on Amnesty (Anne_Baler on Amnesty.org, condition "pre1"), and half saw a typo in the identifier (@AnneBaller on Twitter.com, condition "pre2"). We decided on this to get as many different perceptions as possible. We assumed the participants would most likely recognize the typo but might make a slip with the line on the business card and accept the incorrect assignment of identifiers.
8. To prepare for a more general, less tech-savvy sample, we rephrased "social authentication" as "proof of identity".
9. We changed the phrasing of the scenario, e.g., the reader was addressed more formally.
10. We made several smaller changes to the study documents and added the quiz section to ensure participants were at least once confronted with edge cases of the scenario.
11. When the scenario time was over, we asked the participants whether they wanted to make any further decisions.

3.3 Results

This section describes the lab study where we wanted to test the changes we made to the UI to prevent mistakes seen in the pilot study.

In general, seven out of 18 participants failed the task by sending the data to at least one journalist. Table 2 shows an overview of all the participants and to whom they sent the data. Most tried all available methods. The UI improvements generally prevented the mistakes observed in the pilot study. Nonetheless, the failure rate was still high. Below, we describe the results in detail.

3.3.1 Reasons for Failure

The scenario is considered a failure if a participant sends files to the impersonator. Seven participants failed the scenario for the following reasons: a) in-band safety number comparison (P8, P9, P10, P6), b) clicking too fast (P3), c) gambling for money (P13), and d) emotional stress (P2). The following paragraphs provide more details on those themes.

In-Band Safety Number Comparison The most common pitfall for participants was anchoring their trust in publicly known, unverifiable information, involving in-band exchanges of safety numbers.

P8 saw mismatches in the AC and asked Amira for her postal address and parts of the safety number. They decided that this was secret enough and sent the data. However, in the SOAP case, P8 stayed safe and decided against Anne because of an incorrect SOAP response.

P9 also saw the mismatches (QR, SOAP, safety number) but did not decide to stop and tried to find a way to communicate securely. They asked Michael why the scan failed. Michael said he reinstalled Signal and suggested sending the current chat's safety number, which was the attacker's and not the one on the business card. P9 agreed, compared, and marked the conversation as verified. After that, they tried to determine whether Amira was actually Amira by asking whether Amira knew them, as they assumed they had met when they exchanged business cards. Amira claimed to remember Alex and sent the chat's safety number within that communication. P9 also asked for the work address on the business card, and Amira reported the correct one. After that, Amira was marked as verified.

P9 saw the SOAP mismatch and asked Anne for a different way to verify her. Anna sent the current chat's safety number, but P9 was not entirely convinced, even though they marked Anna as verified. They noticed that some SOAP requests for social media profiles were still unanswered. At that point, P9 ran out of time and told the researcher their next step would be to send the material to Amira and Michael.

P10 was rather insecure and initially seemed overwhelmed by the scenario. They initially wanted to look at the data they

ID	Study Cond.	Sent to			Anne's IdPs				Study length	ATI [4]	Method attempted w/ journalists			Reason to Fail
		Anne	Amira	Michael	A	F	G	T			SOAP	QR	Safety	
P1	pre1	○	○	○	○	○	○	○	42 mins	3.9	-	M	Am	-
P2*†	ctrl	●	●	●	○	○	○	○	51 mins	3.4	-	M	-	Stress
P3†	pre2	●	○	○	○	○	○	○	36 mins	2.3	Am,M	-	-	Fast Clicking
P4	ctrl	○	○	○	●	●	○	●	44 mins	5.1	A	M	Am	-
P5	pre1	○	○	○	●	●	○	●	38 mins	4.3	A, Am	M	Am	-
P6†	ctrl	○	●	●	●	●	○	●	60 mins	3.4	A	M	Am,M	In-Band Comparison
P7	pre2	○	○	○	●	●	●	●	57 mins	4.0	A,Am,M	M	Am	-
P8†	ctrl	○	●	○	●	●	●	●	51 mins	3.0	A,Am,M	M	Am	In-Band Comparison
P9†	pre1	○	●	●	●	○	●	●	45 mins	5.6	A	M	A,Am,M	In-Band Comparison
P10†	ctrl	○	●	●	○	○	○	○	67 mins	3.2	-	M	Am,M	In-Band Comparison
P11	pre2	○	○	○	●	●	●	●	45 mins	3.1	A	M	Am	-
P12	ctrl	○	○	○	●	●	●	●	34 mins	3.8	A,Am	M	Am	-
P13†	pre1	○	●	○	●	○	○	●	52 mins	3.4	A	M	Am	Gambling
P14	ctrl	○	○	○	●	●	○	●	41 mins	5.2	A	M	A,Am,M	-
P15	pre2	○	○	○	●	●	○	○	42 mins	2.3	A	M	Am	-
P16	ctrl	○	○	○	●	●	○	●	36 mins	4.3	A,Am,M	M	Am	-
P17	pre1	○	○	○	○	●	●	●	47 mins	3.1	A	M	Am	-
P18	ctrl	○	○	○	○	●	○	●	32 mins	2.3	A,Am	M	Am	-

Table 2: Overview of the lab study participants' scenario results. Each "●" represents that the participant did what is depicted in the column, e.g., sent the data. The column "Anne's IdPs" marks which IdPs were requested by the participants. The names of providers and journalists are abbreviated (**A**mnesty, **F**acebook, **G**mail, **T**witter, **A**nne, **Am**ira, **M**ichael). The ∗ marks the participant who only continued the scenario after the researcher intervened. † marks participants who failed the scenario. The horizontal line after P11 marks the point where the attacker got stronger (see Section 3.2.1). More details such as the reasons for failure are discussed in Section 3.3.1.

received from Hannah, but as they had never used Android before, they got lost in the data management and needed help from the researcher to go back to Signal. They were told again that they did not need to look at the data for the scenario. They did not know what they should compare for Amira but managed to scan the QR code for Michael and recognize that this failed. Still, they told Michael they had sensible data and asked whether he could verify himself. Michael answered with the chat's safety number. At first, P10 was not sure how to compare the numbers but, after a while, realized that the sent number matched the security number in the settings. Afterward, the same happens with Amira. Anna was also asked for verification, but the scenario time was up.

P6 saw the mismatch for the QR code, safety number, and SOAP after requesting them. They asked Amira and Michael why the numbers were not correct. Both sent the current chat's number, and both received the data afterward. P6 told Anne that Signal said it was not secure to communicate based on the failed SOAP text. They even sent a screenshot when Anne said that this was not the case for her but did not send the data.

P3: Clicking Before Reading P3 was in the pre-registered SOAP condition. They saw a SOAP response without a request when they started a new chat with Anne. They clicked on "Mark Anne as verified" and sent her the data without recognizing this action because the chat was marked green and shown as verified. This makes P3 the only participant whose failure of the scenario is directly attributable to SOAP. After the interview, in which they stated that Anne had already been verified, they were presented with the video and were surprised that they had actually clicked a button. They sent a SOAP request to Amira and Michael after seeing that they had to send a message for the other methods to work. They saw the faulty responses and deleted the chats afterward.

P13: Gambling for Money P13 was not sure who to send or not send the data to. After the time was up, they gambled and sent the data to Amira in the hope of getting more money. Although this is clearly related to the study design, we think it highlights that it was not clear to the participants what the secure and correct way to behave in this scenario was. On a similar note, another participant mentioned during the scenario and the interview that they thought it was strange that all the journalists were unsafe to send data to. They compared it to an exam situation where it seemed strange that all the questions had the same answer. Nevertheless, this participant behaved correctly. It appeared to require some effort for some participants to break off communication.

P2: Emotional Stress P2 initially decided not to contact any journalists, fearing that even sending a message would be too much. After the researcher intervened to tell them it would be all right, P2 went further. They saw the QR code mismatch and decided against SOAP requests, as they assumed they had to send an email and ended up confused. The participant read through the FAQs for safety numbers and ultimately decided to mark every journalist as verified, although they expressed being unsure of whether that was correct. Afterward, P2 sent the data. The participant was clearly highly emotional and insecure at that stage. In the interview, the participant expressed

frustration with their decision but stated they were emotional in the situation and could not think clearly. They were not aware of the safety number printed on the business card of Amira.

3.3.2 Study Conditions: Pre-Registered SOAP

We identified only one case where the condition negatively affected the results. Pre-registered SOAP failed once as P3 auto-clicked the decision. We think an obstacle, e.g., a time restriction or a different visualization, could have prevented that. Only two participants decided to send Anne the data, and they covered both conditions. Seven out of nine participants (pre-reg) sent another SOAP request. Although five of the seven clearly indicated being unsure or seeing the discrepancy. Only one participant did not communicate further with Anne.

3.3.3 Quantitative Data - Perception

It is difficult to compare the results with those from Vaziripour et al. [24] due to different scenarios, methods, and UIs. Nonetheless, with only their and our studies about SA available, we think it is sensible to point out similarities and differences between them. The scenario tested in our study did not involve any direct personal human contact other than through the chat. This was different in the study by Vaziripour et al. [24]. For example, participants could call each other for verification and meet in person to scan the QR code. The researchers found that their proposal of SA ranked higher than the other available methods in the Single-ease-question (SEQ). Conversely, we observed that the tested implementation of SOAP ranked lower than the other two methods provided (see Figure 2b in the Appendix). Potentially in relation to the other available methods, SA ranked much lower in Vaziripour et al.'s study in the trust score than the other extremely high-ranking methods. We observed a mix of perceptions (see Figure 2c). The participants in our study generally trusted all the methods less than the participants in Vazirpour et al.'s study [24]. However, SA still ranked the lowest in both studies. We also asked participants whether they were confident in their decision with the method and saw that SA ranked third in this category while only leading to one failure in the scenario.

3.3.4 Participants' Perceptions and Understandings

We briefly interviewed each participant, asking them about their understanding of the ACs. We found that only two participants had a detailed understanding of safety numbers and the QR code. They used terms like "E2EE" and "public/private keys". One of them studied computer science, where they learned about this, and the other person recognized parallels from email encryption. Four other participants mentioned terms like "E2EE" but had no further concepts of it. They have heard the terms before and connected them to Signal. Although their technical knowledge was limited, many participants conducted the ACs correctly. All participants understood that the QR code, safety number, or accounts should have matched with what was given. All four participants who asked for the safety number via chat detected a mismatch in the QR code or safety numbers beforehand. Only P10 did not see the safety number on the business card. The four participants asked for the number in the chat as a mitigation tactic. All of them were convinced that it is safe to send after receiving the current chat's safety number (see "in-band comparison" in Section 3.3.1).

SOAP, or "proof of identity" as it was called in the study, was known to no one. Speculations on how SOAP worked were, similar to the other ACs, very vague. Often, the participants only stated how they used it and that the accounts should have matched. The participants believed there must be some kind of connection between the accounts provided and the Signal account. It was speculated that this could be based on a one-time token that must be entered (similar to SMS codes that are sent if a phone number is used as an account name), that a person needs to add the number to the account at the IdP, or more generally that the journalists need to log in into the account and do something. Another belief we encountered was that SOAP was based on a setup that happened during the Signal account creation. No participant mentioned safety numbers in their ideas about SOAP.

The participants thought that if the journalists had to take action to create a response, a typo in the account name could occur. Therefore, the participants requested SOAP multiple times to rule out such cases, just as they scanned the QR code multiple times.

According to Section 3.3.3, the participants trusted SOAP less than the other methods. In the interviews, one participant was confused that SOAP gave a valid response, although the QR code was mismatched. Also, the participants thought that account names could somehow be faked or an IdP could be hacked. We additionally found that account names were perceived as private and that the participants were unsure about the processes occurring on the side of the to-be-verified person.

4 Discussion

We conducted a lab study with 18 participants to observe social authentication (SA), an authentication ceremony (AC), in a no-win attack scenario. In this section, we discuss our results from the perspective of our research questions.

4.1 RQ1 & RQ2: Resistance Against Impersonation Attacks - Detection and Reaction

This study observed a SA ceremony in an attack scenario. We were interested in how resistant SA would be against impersonation attacks. So, as a first step, we researched an

attacker who used typo squatting attacks to impersonate the communication partners. While with the simple interface in the pilot study, three participants failed because of SOAP, only one participant in the pre-registered condition failed in the lab study. The UI heavily supported the participants in detecting mismatches when an identifier was given. We applied a strong, opinionated design, e.g., interpreting anything other than the requested identities as incorrect and reducing possible providers to a fixed list. That seemed to help, but we could not measure long-term effects in our setting.

The participants often tried one method, then tried another, changed the journalist, returned to the first, and sometimes retried a previous method. The participants' flow through the tasks was not linear. Not all the participants reacted as hoped to an incorrect SOAP response. Some of the **participants retried** SOAP after seeing an incorrect response or even tried further and asked via text for a way to authenticate the other person. While we found plausible reasons for the first case, we cannot rule out that it is a study artifact. We think this should be investigated further in future studies and considered when designing studies and interfaces. The participants not aborting the communication does not necessarily reflect the hope connected to SA: an intuitive method for recognizing whether you are communicating with the right person. It is also likely that the lab setting influenced the participants, e.g., through demand effects. We, therefore, understand the results as an upper bound for failures in ACs.

The participants without any technical knowledge about what happened **concluded that something was wrong**, although they did not necessarily attribute this to a malicious actor, despite the explicit mention of them in the scenario. Inputting the identifier beforehand helped the automatic detection and, therefore, the automated decision. Based on our data, we do not know whether this can be expected in a real scenario. It is, e.g., unclear where users would source the identifiers from. Anecdotally, in the pilot study, a participant was unsure whether there were unique Facebook identifiers and where to find them. There are paths to help the user here, e.g., if the person's identifier is not known, external means can verify it afterward (e.g., seeing connections in a social graph or validation through a third site). We think there are many possibilities for how this can develop over time, and it is an important area for future work.

While we think what we observed is promising, the sample was too small to draw strong conclusions regarding resistance against impersonation attacks in the real world.

Safety Numbers, on the other hand, did not seem to be resistant to impersonation attacks. While safety numbers were not the focus of the study, we want to highlight that some of the participants failed the scenario because they did an in-band exchange and comparison of key material. As safety numbers comparison is a currently available AC, this should be researched further, as well as whether this has a negative real-world impact. We suggest seeing whether some preventive action can be taken on the client side, e.g., by pattern matching and informing users when they attempt to exchange safety numbers via chat.

4.2 RQ3: Perception and the Role of the Identity Provider

Although only one participant made a mistake with SOAP, the participants were not as confident about their decisions with SOAP as with the other methods. The same trend existed for the perceived usability or trustworthiness of the method to verify their contact. However, the small failure rates contradict that perception. We argue this could be a positive situation for SA. The usability aspect seems solvable, and the participants behaved as intended. But, for the other methods, they behaved insecurely but felt as confident as with SA, creating an "illusion of security" [5]. Regarding SOAP, the participants behaved as hoped. Now, we need to improve the participants' confidence in their own judgment based on SOAP. We are not sure where the difference in the perception of the methods comes from. The sample was too small to make any sensible statistical inferences, but we think further research could investigate the phenomenon.

4.3 Further Observations

This section covers themes beyond our research questions that may offer relevant insights to researchers and practitioners.

Identification of the Person vs. Authentication of the Connection Similar to other studies [3], we observed that the participants did not fully understand how encryption works and, following this, what an attack would look like. We observed, e.g., the assumption that if you have the correct phone number, you will end up with the correct person. In combination with the theme of the "almighty hacker"(see Dechand et al. [3]), participants assumed there is nothing a user can do to protect their communication effectively. So, explaining to the participant that doing something is necessary to communicate with the correct person may be easier than explaining that something is necessary to prevent others from listening. In short, the mental models of Signal's functions did not seem to align enough with the technical reality to understand an attacker. Considering this, it is understandable why the participants fell back to using addresses and shared secrets, or something perceived as such, to identify the other person.

4.4 Protocol/UI Challenges

Multiple requests are not a problem for the protocol per se but can complicate the UI. When designing a protocol and the corresponding interface, designers should remember that the interaction may involve multiple, sometimes canceled, requests. For example, on the one hand, we wanted to ensure that no data were sent with a failed request, but on the other

hand, a typo in the expected identifier was possible and needed to be traceable (false-negative). The participants wanted to believe the other person was legitimate. They were looking for a way to send the data rather than a reason not to send it.

To get the safety number or to receive a SOAP response, a chat contact has to **communicate with the other party**. Depending on the scenario, this communication can be problematic, and participants may hesitate to communicate. If the server is trustworthy, one can reduce the friction here. However, if one also does not trust the server, this is still a problem to be solved. For future studies, communication can be explicitly allowed in the scenario to reduce participants' confusion. In the study, pre-registration caused one failure but helped participants identify issues in other cases.

Vaziripour et al. [24] concluded that the necessary infrastructure for SA "needs to be more trusted than social media companies". We observed that the participants wanted to ask for the journalists' working email addresses. Such **custom providers** are not intended by the (SOAP) protocol. Allowing custom providers also allowed typo attacks on the provider level, making everything even more complicated. It is necessary to determine whether the usage of SA can be reduced to a fixed set of providers, depending on the use case. For example, in a company, setting the list of providers could vary vastly from that of instant messaging for personal use. We suggest finding a way to allow additional, possibly ad-hoc selected providers without impacting security. With SOAP, Linker et al. [11] proposed an interactive communicative way to verify a person. Vaziripour et al. [24] proposed an asynchronous interaction with the previously made public public key. We simulated this in the pre-registered condition after we observed that participants hesitated to even write a single message before verifying a person. We think this hesitation will not appear in most scenarios, but for those where it matters, pre-registration solves a problem. We thus suggest investigating further how an asynchronous solution could be achieved or how the interactive solution can reduce friction.

4.5 Signal Specifics

Some observations made are highly specific to the Signal app and may spark discussions about the UI. Some participants were confused by the way the safety numbers were presented. If a user opens the safety number page, the numbers appear in an animation, giving the impression they are generated just then and would change every time the site is visited.

Signal has the option to use the camera from the start screen. The participants tried using this feature to scan the QR code of the safety number. Here, direct feedback that something was done incorrectly or what type of data might have been scanned could have helped the participants. Signal allows safety numbers to be compared from the clipboard, but no participant was aware of that. When something that looks like a safety number appears within a chat, Signal can provide additional information, e.g., to prevent in-band comparison, but also enhance the sharing of fingerprints between already verified contacts. Similarly to Shirvanian et al. [18], we observed situations where the participants had to compare long numbers across multiple views, but that was not intended and insecure to do in our scenario.

5 Limitations

Conducting a lab study comes with limitations. Participants may behave differently than they would in real life. For our study, this could have led to more interaction and attempts even if the participants thought they should stop. The used reward and risk system is not the same as being a whistleblower and getting caught by the government. But unlike previous studies, which had no risk, we offered a real tradeoff. However, it is still a role play, and we are unaware of the extent of the impact. The setting of a lab study might lead participants to continue because they think there must be a way. We ensured that participants knew that all the connections were potentially insecure and that no communication was also an option. Also, within the scenario, trying different methods to authenticate the journalists was unproblematic. Nonetheless, feedback suggested that the participants liked the scenario and tried to empathize with the situation. We had to pick a fixed set of providers. To not only rely on U.S.A.-based providers, we added Amnesty.org, although it does not offer an identity service to work with SOAP. We do not believe any participant knew this technical detail. Due to a bug with Amira, some of the participants did not need to send a message to get the safety number. We saw that sending a message made the participants hesitant, but ultimately, they all decided this was not a show-stopper.

6 Conclusion

To test a new social authentication (SA) protocol called SOAP and compare it with traditional ACs, we developed a scenario-based lab study where participants take over the role of whistleblowers and try to gauge whether the connections to journalists contacted via the Signal app are secure. Based on a pilot study, we improved an interface for SOAP and made it similarly usable as manual safety number comparison or QR codes. We found that although the participants did not know how SA worked, they behaved mostly securely, and mistakes were more often made in existing ACs. These findings make us optimistic about SA as a usable AC. While our sample size was rather small, and our scenario may not directly translate to a realistic real-world situation (e.g., at the current time, we do not recommend whistleblowers to use SOAP), it provided the participants with understandable reasoning and motivated them to act securely. With the study design, we provide a template for further research and comparison of ACs.

7 Acknowledgements

We are grateful to our shepherd and the anonymous reviewers for their valuable comments and suggestions. We thank the Werner Siemens-Stiftung (WSS) for their generous support of this project.

References

[1] Whistleblower Aid. Become a whistleblower. https://whistlebloweraid.org/become-a-whistleblower/signal/. Accessed: 30 October 2023.

[2] Mashari Alatawi and Nitesh Saxena. SoK: An Analysis of End-to-End Encryption and Authentication Ceremonies in Secure Messaging Systems. In *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, WiSec '23, page 187–201, New York, NY, USA, 2023. Association for Computing Machinery.

[3] Sergej Dechand, Alena Naiakshina, Anastasia Danilova, and Matthew Smith. In Encryption We Don't Trust: The Effect of End-to-End Encryption to the Masses on User Perception. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 401–415, June 2019.

[4] Thomas Franke, Christiane Attig, and Daniel Wessel. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction*, 35(6):456–467, April 2019.

[5] Amir Herzberg and Hemi Leibowitz. Can Johnny finally encrypt?: Evaluating E2E-encryption in popular IM applications. In *Proceedings of the 6th Workshop on Socio-Technical Aspects in Security and Trust*, pages 17–28, Los Angeles California, December 2016. ACM.

[6] Amir Herzberg, Hemi Leibowitz, Kent Seamons, Elham Vaziripour, Justin Wu, and Daniel Zappala. Secure Messaging Authentication Ceremonies Are Broken. *IEEE Security & Privacy*, 19(2):29–37, March 2021.

[7] Sakshi Jain, Neil Zhenqiang Gong, Sreya Basuroy, Juan Lang, Dawn Song, and Prateek Mittal. New Directions in Social Authentication. In *Proceedings 2015 Workshop on Usable Security*, San Diego, CA, 2015. Internet Society.

[8] Keybase. Keybase book. https://book.keybase.io/docs/server#meet-your-sigchain-and-everyone-elses. Accessed: 8 January 2024.

[9] Sean Lawlor Lewi, Kevin. Deploying key transparency at WhatsApp. https://engineering.fb.com/2023/04/13/security/whatsapp-key-transparency/, April 2023.

[10] Guardian News & Media Limited. How to contact the guardian securely. https://www.theguardian.com/help/ng-interactive/2017/mar/17/contact-the-guardian-securely. Accessed: 30 October 2023.

[11] Felix Linker and David Basin. Soap: A social authentication protocol. https://arxiv.org/abs/2402.03199, 2024.

[12] Lee Livsey, Helen Petrie, Siamak F. Shahandashti, and Aidan Fray. Performance and Usability of Visual and Verbal Verification of Word-Based Key Fingerprints. In Steven Furnell and Nathan Clarke, editors, *Human Aspects of Information Security and Assurance*, volume 613, pages 199–210. Springer International Publishing, Cham, 2021. Series Title: IFIP Advances in Information and Communication Technology.

[13] Marcela S. Melara, Aaron Blankstein, Joseph Bonneau, Edward W. Felten, and Michael J. Freedman. CONIKS: Bringing key transparency to end users. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 383–398, Washington, D.C., August 2015. USENIX Association.

[14] moxie0. Safety number updates. https://signal.org/blog/safety-number-updates/. Accessed: 2 January 2024.

[15] ZEIT ONLINE. Appell an potenzielle whistleblower. https://www.zeit.de/administratives/2019-01/technologiebranche-whistleblower-suche. Accessed: 30 October 2023.

[16] The Washington Post. Submit an anonymous news tip. https://www.washingtonpost.com/anonymous-news-tips/. Accessed: 30 October 2023.

[17] Svenja Schröder, Markus Huber, David Wind, and Christoph Rottermanner. When SIGNAL hits the Fan: On the Usability and Security of State-of-the-Art Secure Mobile Messaging. In *Proceedings 1st European Workshop on Usable Security*, Darmstadt, Germany, 2016. Internet Society.

[18] Maliheh Shirvanian, Nitesh Saxena, and Jesvin James George. On the Pitfalls of End-to-End Encrypted Communications: A Study of Remote Key-Fingerprint Verification. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, ACSAC '17, pages

499–511, New York, NY, USA, December 2017. Association for Computing Machinery.

[19] Signal. Signal messenger: Speak freely. https://signal.org/. Accessed: 13 February 2024.

[20] Joshua Tan, Lujo Bauer, Joseph Bonneau, Lorrie Faith Cranor, Jeremy Thomas, and Blase Ur. Can Unicorns Help Users Compare Crypto Key Fingerprints? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3787–3798, Denver Colorado USA, May 2017. ACM.

[21] The Telegram Team. Colorful calls, thanos snap effect, and an epic update for bots. https://telegram.org/blog/calls-and-bots. Accessed: 2 January 2024.

[22] Telegram. Faq for the technically inclined. https://core.telegram.org/techfaq#man-in-the-middle-attacks. Accessed: 8 January 2024.

[23] The New York Times. Got a confidential news tip? https://www.nytimes.com/tips. Accessed: 30 October 2023.

[24] Elham Vaziripour, Devon Howard, Jake Tyler, Mark O'Neill, Justin Wu, Kent Seamons, and Daniel Zappala. I Don't Even Have to Bother Them! Using Social Media to Automate the Authentication Ceremony in Secure Messaging. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, New York, NY, USA, May 2019. Association for Computing Machinery.

[25] 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 *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 47–62, Baltimore, MD, August 2018. USENIX Association.

[26] Elham Vaziripour, Justin Wu, Mark O'Neill, 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 *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 29–47, Santa Clara, CA, July 2017. USENIX Association.

[27] Justin Wu, Cyrus Gattrell, Devon Howard, Jake Tyler, Elham Vaziripour, Daniel Zappala, and Kent Seamons. "Something isn't secure, but I'm not sure how that translates into a problem": Promoting autonomy by designing for understanding in Signal. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 137–153, 2019.

[28] Tarun Kumar Yadav, Devashish Gosain, Amir Herzberg, Daniel Zappala, and Kent Seamons. Automatic detection of fake key attacks in secure messaging. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, page 3019–3032, New York, NY, USA, 2022. Association for Computing Machinery.

[29] Süddeutsche Zeitung. So erreichen sie das investigativ-team der süddeutschen zeitung. https://www.sueddeutsche.de/projekte/kontakt/#messenger. Accessed: 30 October 2023.

A Appendix

The original study material in German can be found online at https://osf.io/dsyfr/. Due to space constraints, we have only included the translated versions in this paper.

A.1 Scenario

This are the translated scenario texts available to the participants, including the payment description for both studies.

A.1.1 Scenario Text for the Pilot Study

The study consists of a role play in which you take on the role of Alex. The scenario is described in the following text. Please read the text carefully and put yourself in the situation.

Scenario card Your name is Alex and you live in a country ruled by an authoritarian regime. Both blanket and targeted surveillance is a daily phenomenon. You work in a high-ranking government agency. A colleague, Hannah, has gained access to extremely sensitive information about high-level corruption and shared it with you in encrypted form through the Signal app a few months ago. This information includes revelations about illegal activities by politicians.

You want this information to be made public. In order to avoid drawing suspicion to Hannah, who had access to the data, you have decided to wait a few months and then send the data to journalists. The time has now come and you can begin.

You already have business cards from three trustworthy investigative journalists from abroad. You have received them personally and you trust the information on them. All journalists are known for their integrity and have already uncovered a number of major scandals. All three journalists offer whistleblowers that they can be contacted securely via the Signal app.

You have a rough idea of how such a contact works: First you send the data to the journalist. The journalist then does research and checks whether the data is genuine. Once they are satisfied, they publish the story. This can take a while. The archive containing the data and explanation can be found in your Signal app in the chat with Hannah. You are familiar with the content and the exact content does not matter for the study.

Your goal is to ensure that all three journalists receive the data about the corruption. Considering the dangers you and Hannah face if your government's intelligence agencies find out that you have leaked the data, it is crucial for you to make sure that you communicate with the journalists in encrypted form using the Signal app. You are sure that as long as you use the Signal app correctly, the secret services will not be powerful enough to break the encryption or access the metadata.

Bonus payment: You currently have 5€ in your account. For every journalist you successfully send the data to, you will receive another €5.

However, if you are caught by the secret service, you will end up in prison and will not receive any payment. So only send the data if you are sure that the Signal app will protect you. You will receive the exam bonus of 2% points even if you end up in prison. If you are not caught, you will receive the 2% points and the money from your account.

Instructions

Signal Signal is an encrypted messenger and phone app. Signal saves your number, but does not create a log file for your incoming or outgoing communication. Signal is easy to use: Open the app and tap the pencil icon (bottom right on Android phones) to write a new message. Enter the desired phone number in the search field. You can now send an encrypted message via Signal.

How do I take screenshots? Press and hold the "On/Off" button and "Volume down" button on your phone at the same time for about one second.

A.1.2 Scenario Text for the Lab Study

The study consists of a role play in which you take on the role of Alex. The scenario is described in the following text. Please read the text carefully and put yourself in the situation.

Scenario Description

Your name is Alex and you live in a country ruled by an authoritarian regime. Both blanket and targeted surveillance happen on a daily basis. You work in a high-ranking government agency. A colleague, Hannah, has gained access to extremely sensitive information about high-level corruption and shared it with you in encrypted form through the Signal app a few months ago. This information includes revelations about illegal activities by politicians. You want this information to be made public. In order to avoid drawing suspicion to Hannah, who had access to the data, you have decided that you will wait a few months and then you (Alex) will send the data to journalists. The time has now come and you can begin.

You already have business cards from three trustworthy investigative journalists from abroad. You have received the business cards personally from the journalists and you trust the information on them. All journalists are known for their integrity and have already uncovered a number of major scandals. All three journalists offer whistleblowers that they can be contacted securely via the Signal app.

You have a rough idea of how such a contact works: First you send the data to the journalist. The journalist then researches and checks whether the data is genuine. If the journalist is convinced, the story is published. This can take a while. The archive containing the data and explanation can be found in your Signal app in the chat with Hannah. They are familiar with the content but the exact content does not matter for the study.

Your goal is for all three journalists to receive the data on corruption. Considering the dangers you and Hannah face if your government's intelligence services find out that you have leaked the data, it is crucial for you to ensure that you communicate with the journalists in encrypted form using the Signal app. You are sure that as long as you use the Signal app correctly, the intelligence services will not be able to break the encryption or access the metadata.

Payment: You will receive a basic payment of €15 after completing the study. You also have the option of receiving a bonus of up to €9.

You should send the data to the journalists with a secure connection - and only to those with a secure connection.

A decision must be made for each of the three journalists individually:
- If the connection is secure, the data must be sent.
- If the connection is insecure, no data may be sent.

For each correct decision you receive a €3 bonus, i.e. up to €9 in total. But: No bonus is awarded, - if data is sent via at least one insecure connection - or if no data is sent although there is at least one secure connection.

How to send messages with Signal

Open the app and tap the pencil icon at the bottom right to write a new message. Enter the desired phone number in the search field. You can now send an encrypted message via Signal.

What is a secure connection?

The Signal app offers you methods to ensure that you are communicating with the right person and correctly encrypted. If you cannot use the app to ensure that the connection is secure, you should assume that the connection is insecure.

A.2 Survey

The survey varied slightly in the pilot and the lab study. Social authentication was called "Proof of identity (Identitätsnachweis)" in the lab study and participants were addressed more formally. The questions that were exclusively part of a study or edited a lot are marked.

Q1: Below are some questions about the methods you interacted with during the study. The lab study includes a role play. However, please do not fill out this questionnaire in the role of Alex, but as yourself. (Type: Text)

Q2: Please enter your study pseudonym (Type: Text Entry)

Q3: Do you use the Signal app independently of the study? (Type: MC) Answer Choices: *"No", "Yes, Rarely", "Yes, Often"*

Q4: Before you took part in the study: For how many of your chat contacts did you use a safety number (e.g. in Whatsapp or Signal) to verify the contact? (Type: MC) Answer Choices: *"With none of my chat contacts", "With some of my chat contacts", "With about half of my chat contacts", "With most of my chat contacts", "With almost all of my chat contacts"*.

Q5: During the study, you tested up to three different methods of verifying a contact via Signal. a) via QR code scan b) comparing safety numbers c) via account affiliation on platforms (social authentication). The following questions are about your thoughts on exactly these methods. (Type: Text)

Q6: Which of the methods did you use in the course of the study? (Type: MC) Answer Choices: *"QR code", "safety number", "social authentication"*

Q7: How much do you agree with the following statement: I have confidence in this method of verifying safety numbers in Signal. *Lab study: I have confidence in this method for verifying the identity of my conversation partners*. (Type: Matrix) Items: *"safety number", "social authentication", "QR code"*
Scale (5): *Strongly disagree, Somewhat disagree, Neither agree nor disagree/neutral, Somewhat agree, Strongly agree*

Q8: How much do you agree with the following statement: I am sure that I made the right decision when using the method. (Type: Matrix) Items: *"safety number", "social authentication", "QR code"*
Scale (5): *Strongly disagree, Somewhat disagree, Neither agree nor disagree/neutral, Somewhat agree, Strongly agree*

Q9: In terms of verifying with the appropriate method: Overall, how difficult or easy was it to complete the task? *Lab study: Related to verifying identity using the appropriate method: How did you find completing the task* (Type: Matrix) Items: *safety number, social authentication, QR code*
Scale (5): *Very difficult, Very easy*

only lab study: Q10: Please mark which method you would choose if you had to verify a friend. (Type: MC) Items: *"safety number", "social authentication", "QR code"*

Q11: Is there anything else you would like to tell us about the methods? (Type: Text Entry)

Q12: Please indicate your level of agreement with the following statements. (Type: Matrix) Items: *"I understood the scenario", "I think the scenario is plausible", "I thought myself into the scenario", "The chance of getting money motivated me to contact as many journalists as possible", "The risk of losing money motivated me to be careful", The financial incentive helped me to empathize with the scenario", "Without a financial incentive I would not have taken the scenario so seriously", "Without a financial incentive I would not have gone to so much trouble to check the security numbers"*.
Scale (5): *Strongly disagree, Somewhat disagree, Neither agree nor disagree/neutral, Somewhat agree, Strongly agree*

Q13: The following is about your interaction with technical systems. By 'technical systems' we mean apps and other software applications as well as complete digital devices (e.g. cell phone, computer, TV, car navigation). Please indicate your level of agreement with the following statements.

(Type: Matrix) Items: *"I like to take a closer look at technical systems", "I like to try out the functions of new technical systems", "I primarily deal with technical systems because I have to", "When I have a new technical system in front of me, I try it out intensively", "I like to spend a lot of time getting to know a new technical system", "It is enough for me that a technical system works, I don't care how or why", "I try to understand exactly how a technical system works", "It is enough for me to know the basic functions of a technical system", "I try to make full use of the possibilities of a technical system"*.
Scale (6): *Not true at all, Not true to a large extent, Rather not true, Rather true, Moderately true, Completely true*

Q14: How old are you? (Type: Text Entry) *only lab study:*

Q15: Which gender do you feel you belong to? (Type: MC) Answer Choices: *"Female", "Male", "Diverse", "I would like to describe myself:", "Not specified"*

Q16: Which employment situation suits you? What in this list applies to you? Please note that gainful employment is understood to mean any paid or income-related activity associated with an income. (Type: MC) Answer Choices: *"Full-time employment", "part-time employment", "partial retirement (regardless of whether in the working or release phase)", "marginally employed, 450-euro job, mini-job", ""one-euro job" (in receipt of unemployment benefit II)", "occasionally or irregularly employed", "In vocational training/apprenticeship", "In retraining", "Voluntary military service", "Federal voluntary service or voluntary social year", "Maternity leave, parental leave, parental leave or other leave of absence (click on the relevant option for partial retirement)", "Not gainfully employed (including: Pupils or students who do not work for money, unemployed, early retirees, pensioners without additional income)"*

Q17: If you are not in full-time or part-time employment: Please say, which group on this list you belong to. (Type: MC) Answer Choices: *"Pupils at a general school", "students", "pensioners, retired, early retired", "unemployed", "permanently disabled", "housewives/househusbands", "other, namely:"*

Q18: Please note that it is important that the questions asked in this questionnaire are answered by each participant independently and without prior knowledge of the study. This ensures the integrity and quality of our data. We therefore ask you not to share any information about the content of the study or the questions of this questionnaire with other people for 2 weeks and your answer to the next question will have no effect on you, your bonus points or bonus payment! But it is very important for us that you answer honestly. Did you already know the details of what happens in the study before participating in the study? *Lab study:*Please note that it is important that the questions asked in this questionnaire are answered by each participant independently and without prior knowledge of the study. This ensures the integrity and quality of our data. We therefore ask you not to share any information about the content of the study or the questions in this questionnaire with other people for one week. Your answer to the next question will not affect you or the money you receive at the end of the study! But it is very important to us that you answer honestly. Before participating in the study, did you already know details about what will happen in the study? (Type: MC) Answer Choices: *"Yes", "No"*

Q19: What did you know and how do you think it affected you? (Type: Text Entry)

Q20: Thank you for completing the questionnaire. Now please turn to the person in the room. (Type: Text)

A.3 Quiz

In the lab study the participants had a to complete a quiz after reading and before starting the scenario. They could answer questions as often until they had all correct.

Q21: The following questions are intended to ensure that you have carefully read and understood the assignment. You can use all available documents to answer the questions.

Q22: What is the name of the person you are supposed to play? (Type: MC)
Answer Choices: "Alex", "Hannah", "Friedrich", "Eva"

Q23: What should you do if you have established a secure connection with a journalist? (Type: MC)
Answer Choices: "Send the data to the journalist and try to contact other journalists", "Cancel the contact", "Let Hannah know", "Send the data to the journalist. The task is then completed."

Q24: What should you do if you cannot ensure that a connection to a journalist is secure? (Type: MC)
Answer Choices: "Cancel the contact", "Send the data to the journalist", "Let Hannah know".

Q25: Under what conditions should you send the data to whom? (Type: MC)
Answer Choices: "All journalists, even if I can't be sure that the connections are secure", "Every journalist with whom there is a secure connection", "Hannah".

Q26: In which situations does the bonus payment increase? (Mehrfachnennung ist möglich.) (Type: MC) *Answer Choices: "A connection is not secure and I am not sending data", "A connection is secure and I am sending data", "A connection is not secure and I am sending data", "A connection is secure and I am not sending data"*.

Q27: What possible situations can occur in the study? (Mehrfachnennung ist möglich.) (Type: MC) *Answer Choices: "All connections are secure and I send the data to all journalists", "No connection is secure and I don't send the data to anyone", "Some connections are secure and I send the data there"*.

Q28: Which statement is true? (Type: MC)
Answer Choices: "The information on the business cards is correct", "The information on the business cards may be incorrect".

A.4 Interview Guideline

1. Why did you decide to act the way you did with the journalists? (Go through it step by step, was impersonation a conceivable option?)
2. How do you think the methods work? (safety number, QR code, social authentication)
3. Do you have an idea where you would like to apply such a method?
4. Which method would you use if you had to verify a friend? (Focus on why)
5. Would you be willing to use your accounts for social authentication?
6. Would you behave differently as a whistleblower outside of the study?

A.5 Debriefing Guideline

1. Have the payment form filled out.
2. Ask the participant not to talk about the study for one week. Explain how things work would render the data unusable.
3. Explain the objectives: To see if an impersonation attack is detected and what the reactions are. We were also interested in what are the thoughts concerning the procedure.
4. Explain: Security numbers must come through a different channel than the conversation. They change if there is an eavesdropper, but also if, for example, one changes their phone and reinstalls Signal.
5. Are there any questions?

A.6 Business cards

This is the information on the business cards participants had available. All the information except the phone numbers were made up.

Amira

- Amira Patel
- Investigative journalist
- Hallentorstraße 4, 20654 Hamburg
- Phone: {removed as the numbers actually exist}
- Mail: amira_patel@newsorg.de
- Signal Safety Number: 72500 10336 57813 26686 75084 04894

Anne

- Anne Baler
- Investigative journalist
- Isarwege 15, 80542
- Phone: {removed as the numbers actually exist}
- Mail: anne_baler@newsunion.de
- Twitter.com: @AnneBaler
- Facebook.com: Anne_Baler
- *(only pilot study)* Instagram.com: @AnneBaler
- *(only pilot study)* LinkedIn: anne-baler-98746524b
- *(only lab study)* Amnesty.org: anne-baler-98746524b

Michael

- Michael Kobel
- Investigative journalist
- Maurerweg 28, 10235 Berlin
- Phone: {removed as the numbers actually exist}
- Mail: michael_kobel@newsbody.de
- {Image of a QR Code}

A.7 Playbook

This is the playbook the researcher, acting as the journalists, used. It was extended when necessary, and the wording varied slightly from case to case.

Participant: Hello
Journalist: Hello|Hi|Good morning|Good afternoon|Good evening|Good night

Participant: Hello, my name is Alex
Journalist: Hi, I'm [Person Name]

Participant: Hi, do you still know me?
Journalist: No, I don't know whose number this is.
Participant: I got your business card some time ago.
Journalist: I give these to a lot of persons. Sorry that I cannot recall you.
Participant: It's me, Alex.
Journalist: Ah, I can briefly remember.

Participant: I cannot verify your safety number on the business card. Do you have another idea?
Journalist: Yes, my new number is [safety number].

Participant: Can you verify yourself?
Journalist: Yes, I can send you my current safety number.

Participant: Your safety number is wrong
Journalist: My number has changed. My new one is [safety number of conversation].

Participant: Your QR Code does not match.
Journalist: I lost my old phone and had to buy a new one.

Participant: Do you have a [safety number|QR Code] for me?
Journalist: [Current safety number|QR Code sent]

Participant: I cannot contact you because Signal says, there is a chance that the conversation is not secure
Journalist: I don't see this message.

Participant: Screenshot of failed SOAP message
Journalist: My accounts were hacked before and I had to create new accounts. I don't have new ones on [missing Providers here]. On [changed Username providers], I have a new user handle.

Participant: How does your [email, address, phone number, other PII] look like?
Journalist: [Publicly available data]

Participant: I would like to send you the data
Journalist: Okay, I'm ready

Participant: [data sent]
Journalist: Thank you! I will check the files with my colleagues and get back to you.

Participant: [SOAP request]
Journalist: [SOAP response]

Participant: Did you arrive safely home?
Journalist: Sorry, do we know each other?

Participant: Have you developed any ideas for our project?
Journalist: What do you mean?

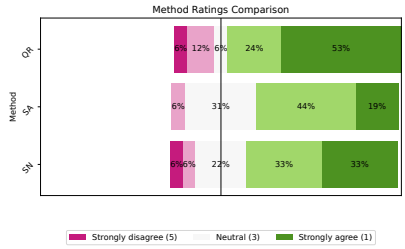
A.8 Additional Tables and Figures

Provider	# of requests
Work email	9
Twitter	8
LinkedIn	8
Instagram	8
Facebook	7
Gmail	4
Reddit	3
Telekom	2
Pinterest	2
iCloud	2

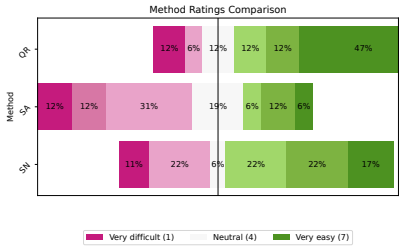
Table 3: Frequency of how often each IdP was requested in the pilot study. The work email is the provider most frequently requested, and in the current protocol proposal, it is not included.

ID	Sent to			ATI [4]	Reason for Failure
	Anne	Amira	Michael		
CS-1	○	○	○	3.9	
CS-2	○	●	○	6.0	Typosquatting
CS-3	○	●	●	5.6	Typosquatting
CS-4	○	○	○	5.8	
CS-5	○	○	○	5.3	
CS-6	●	●	○	4.4	Typosquatting
CS-7	●	●	●	2.9	Marking
CS-8	○	○	○	2.9	
CS-9	○	○	○	4.4	

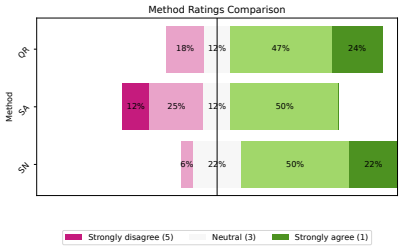
Table 4: Overview of the results of the pilot study participants. Four sent data to at least one journalist (marked by ●). The "Reason for Failure" column matches a theme in Section 3.2.6.



(a) How confident were participants with their decision? (Q8)

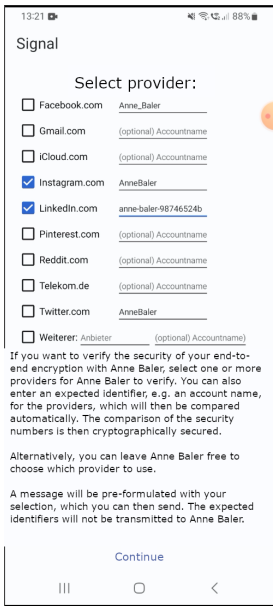


(b) Single-ease-question (SEQ) ratings of the methods. (Q9)

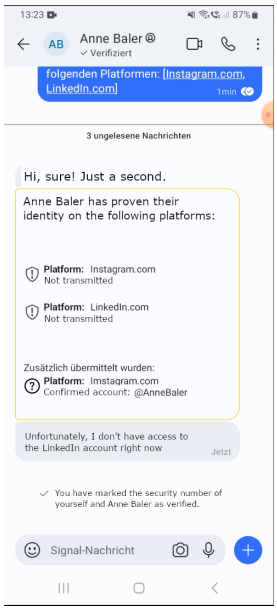


(c) How much do the participants trust the method? (Q7)

Figure 2: The ratings of the methods by the participants of the lab studies. The "n"s differ slightly because not each participant used all the methods. "SN" is short for "safety number" and "SA" is short for "social authentication".

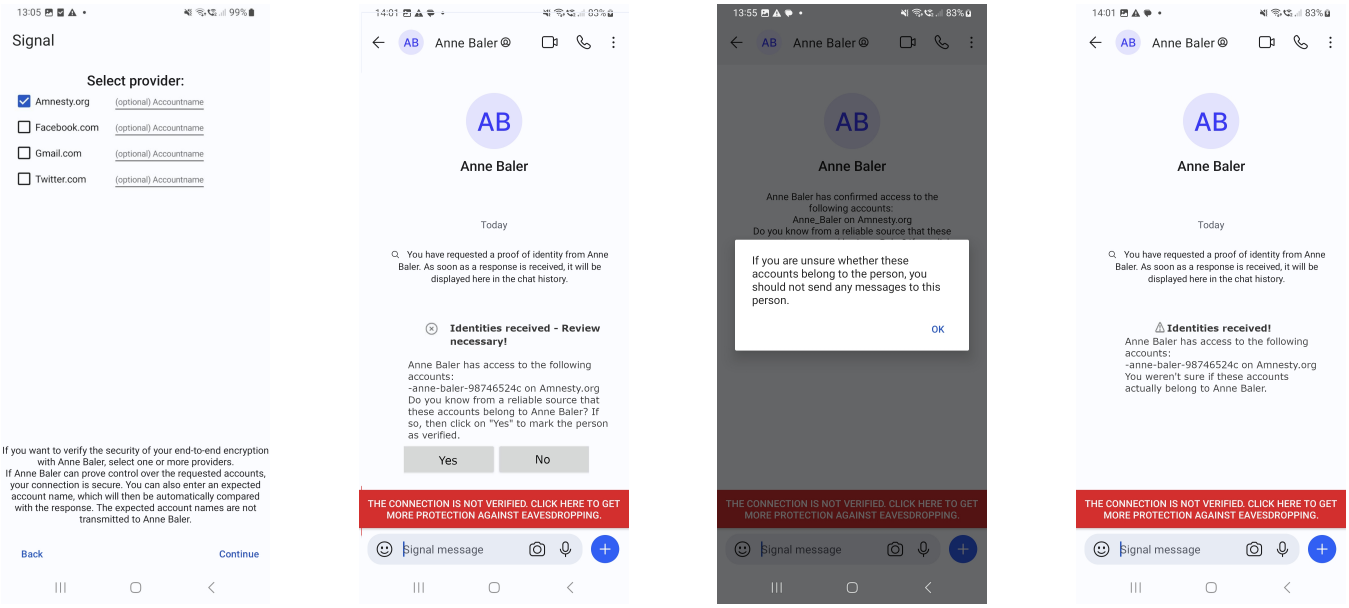


(a) The participant requested two proofs from two providers and filled in the identities.



(b) The response was incorrect due to a typo in one IdP and another identifier not being transmitted; however, the participant did not notice the typo and incorrectly marked Anne as verified.

Figure 3: Translated screenshots of the SOAP request flow from P10 (pilot study).



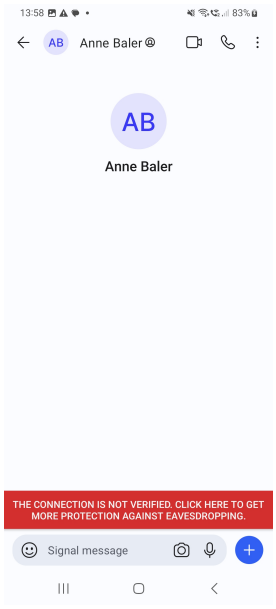
(a) A user can request proof for an account on an IdP without any identity.

(b) As it cannot be automatically decided whether the identifier is correct, the user must make a decision.

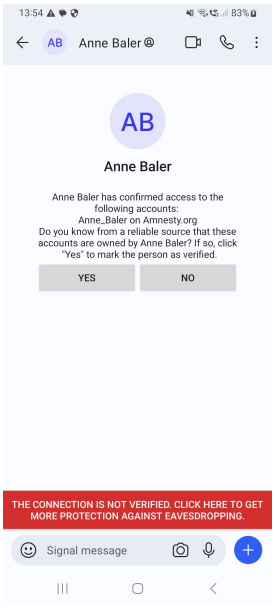
(c) In combination with the decision, the user is informed about possible consequences.

(d) And later reminded what they decided.

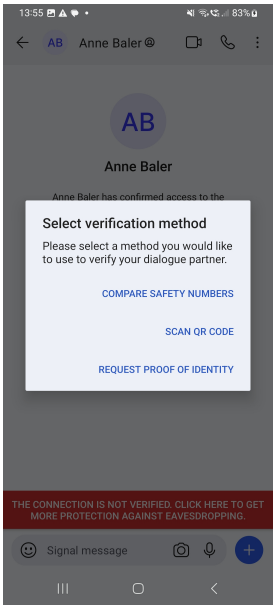
Figure 4: Translated screenshots of a SOAP (lab study version) request flow without identifiers.



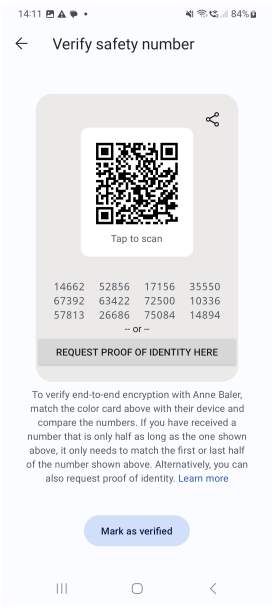
(a) The screenshot shows what a fresh chat looks like if the person was previously added as a contact. The red button nudged the participant to click on it and find the ceremonies.



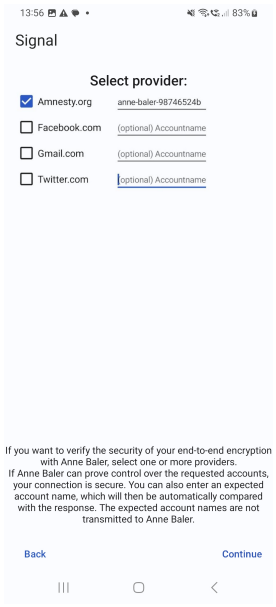
(b) If the participant was in the pre-registration condition, they saw a non-requested SOAP answer.



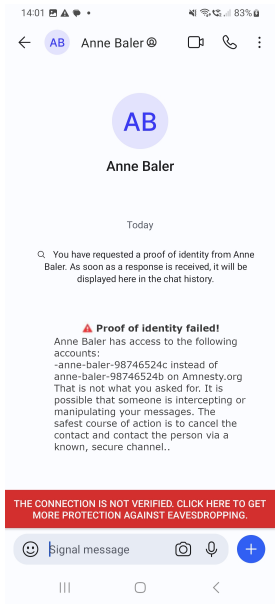
(c) The screenshot shows the menu that led to the ceremonies. Clicking on the QR code and safety number opened the existing site, just slightly modified (see.5d).



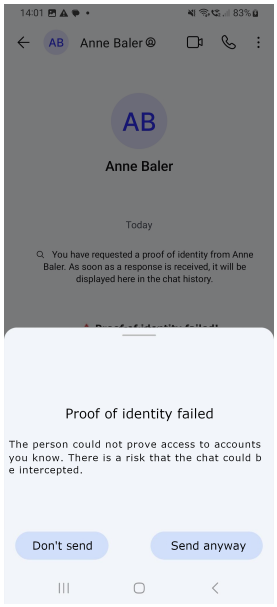
(d) The screenshot shows the menu page where the user can see the safety number, the QR code, and request SOAP. The information text was slightly adapted so that users knew which part of the safety number to compare, and the prompt for social authentication was added.



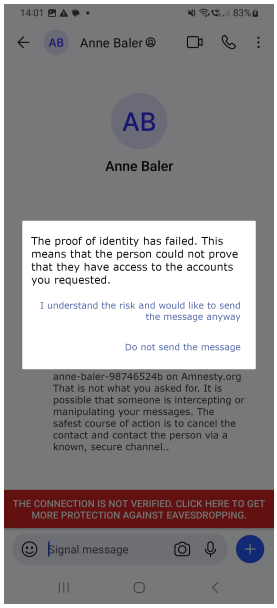
(e) The screenshot shows how a SOAP request code could be made. Selecting an IdP and an identifier. By clicking on next, a SOAP request is sent. The expected identifier is not sent to the recipient.



(f) The SOAP request is sent automatically, and the user is informed in the chat. When the user receives a SOAP response, it is parsed, and compared against the expected identifiers and IdPs. This SOAP response is incorrect, as the identifier is not as expected. The user is informed of that.



(g) If a user wants to send a message to a contact where a SOAP response was incorrect, the app warns the user, similar to when the safety number of a verified contact changes.



(h) To send a message, the user has to click two additional times.

Figure 5: Translated screenshots of an example SOAP (lab study version) request flow with expected identifiers filled in.