Cryptographic Deniability: A Multi-perspective Study of User Perceptions and Expectations

Tarun Kumar Yadav, *Brigham Young University;* Devashish Gosain, *KU Leuven;* Kent Seamons, *Brigham Young University*

https://www.usenix.org/conference/usenixsecurity23/presentation/yadav

This paper is included in the Proceedings of the 32nd USENIX Security Symposium.

August 9–11, 2023 • Anaheim, CA, USA

978-1-939133-37-3

Open access to the Proceedings of the 32nd USENIX Security Symposium is sponsored by USENIX.

Cryptographic Deniability: A Multi-perspective Study of User Perceptions and Expectations

Tarun Kumar Yadav
Brigham Young University

Devashish Gosain
KU Leuven

Kent Seamons
Brigham Young University

Abstract

Cryptographic deniability allows a sender to deny authoring a message. However, it requires social and legal acceptance to be effective. Although popular secure messaging apps support deniability, security experts are divided on whether it should be the default property for these applications. This paper presents a multi-perspective, multi-methods study of user perceptions and expectations of deniability. The methodology includes **(1)** qualitative analysis of expert opinions obtained from a public forum on deniability, **(2)** qualitative analysis of semi-structured interviews of US participants, **(3)** quantitative analysis of a survey (**n=664**) of US participants, and **(4)** qualitative and quantitative analysis of US court cases with help from a legal expert to understand the legal standpoint of deniability. The results show that deniability is not socially accepted, and most users prefer non-repudiation. We found no US court cases involving WhatApp that consider deniability. Significant human-centered research is needed before deniability can adequately protect vulnerable users.

1 Introduction

Cryptographic deniability allows the sender of a message to deny they sent it with no cryptographic evidence to refute their claim. Deniability[1] mimics face-to-face communication, where a person can later deny they said something. Off-the-record (OTR) messaging [2] first introduced cryptographic deniability to instant messaging, and Signal and WhatsApp provide it by default.

Deniability on popular messaging apps seems promising; users can freely communicate their thoughts and later deny any leaks that have potentially adverse consequences. But on the other hand, deniability is a liability for users that need to hold a message sender accountable. In a recent controversial case in Romania, a celebrity was arrested on allegations of human trafficking and rape [14]. The court is investigating WhatsApp chats in which the person allegedly lured the woman into trafficking. It is imperative for the woman that the messaging app provides non-repudiation. In contrast, deniability is crucial for the suspect.

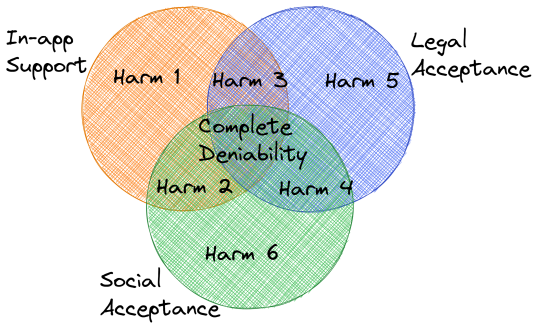


Figure 1: Venn diagram showing the relationship between application support, social acceptance, and legal acceptance for deniability. If any is missing, users are at risk of harm.

This case shows how a sender and recipient may want contradictory properties (deniability and non-repudiation[2]) on the same message. However, providing both properties simultaneously for the same message is infeasible. Even among security experts, there is no consensus on whether deniability should exist as the default property for Internet communication [9] (*e.g.,* over instant messaging apps). Thus, we must study users' needs and preferences as we chart a path forward.

Deniability *requires* social and legal acceptance to be effective. Senders unaware of whether a system supports deniability will be unable to use it. Moreover, users aware that an app supports deniability may have a false sense of security if they do not understand deniability's social or legal acceptance.

There are three necessary components that, synergistically, should make deniability practical and harmless: (1) in-app

[1] In this paper, we use cryptographic deniability and deniability interchangeably.

[2] Non-repudiation assures the integrity and origin of data in such a way that the integrity and origin can be verified and validated by a third party as having originated from a specific entity in possession of the private key [15].

support of deniability, (2) social acceptance of deniability, and (3) legal acceptance. In-app support means that all the messages generated from the IM apps are cryptographically deniable. Social acceptance means users think that messages in the app are deniable and, therefore, will assume they cannot use messages as evidence in society or court. Finally, legal acceptance means that the courts accept deniability and may not consider messages as evidence.

As shown in Figure 1, the following three harms are possible if an app supports deniability but lacks social or legal acceptance. *Harm 1:* deniability is not accepted socially or legally, so an attacker can forge a message that will be trusted socially and legally. *Harm 2:* deniability is accepted socially but not legally, so users assume they can deny a message in court when they cannot. *Harm 3:* deniability is accepted legally but not socially, so users assume a message can be used as evidence in court when it cannot.

Similarly, *Harms* 4, 5, and 6 are possible if an app does not support deniability but has social or legal acceptance. For example, if deniability is accepted socially but not legally, users may assume they can always deny a message, but the courts will accept it as evidence.

We conducted a *mixed method multi-perspective* study on users' perceptions and expectations of deniability for their online communication. Given the limited research on deniability from a user's perspective, we began with a **(1)** qualitative analysis of expert opinions obtained from a public forum [9] where experts discussed the advantages and challenges of deniability. From this, we identified the following four research questions:

RQ1: *What is the social acceptability of denying an actual WhatsApp chat by a sender to a third party? How different is it from denying oral communication, and why?*

RQ2: *What is the user's understanding of deniability in secure messaging after reading the standard definition from the OTR home page?*

RQ3: *What authentication properties (i.e., deniability, non-repudiation, anonymity) do users want across various Internet applications, and why?*

RQ4: *How credible is a WhatsApp chat as evidence in a legal setting?*

To address RQ 1–3, we **(2)** completed a qualitative analysis of semi-structured interviews to identify various themes across these topics (*e.g.,* a user as recipient denying the communication). To quantify our observations about the identified themes on a larger scale, we **(3)** performed a quantitative analysis of a survey (**n=664**) of US participants. Finally, to explore the legal acceptance of deniability (RQ4), we **(4)** performed a qualitative and quantitative analysis of US court cases where WhatsApp chats were potentially used as evidence.

The key takeaways from our research are:

- **Deniability is not socially accepted** If deniability is socially accepted, the users' trust in oral and in-app claims should ideally be the same. But our analysis reveals that users trust in-app chats significantly more than oral claims from a participant in a conversation. The trust depends on the claimant's relationship with the user and the other party in the conversation. The most significant difference between oral vs. in-app trust was for an untrusted claimant, which makes users vulnerable to social engineering attacks that leverage deniability.

- **OTR's deniability definition leads to a false sense of security or a lack of trust in an application** Only 0.6% of participants interpreted OTR's deniability definition accurately. Around 64.8% of users thought they understood the definition but did not. 32% of participants believed that the deniability definition is self-contradictory.

- **Most users prefer non-repudiation for their Internet communications, including messaging apps** When asked what property they want, 60.2% of users desire only non-repudiation, whereas 12.7% and 4.5% desire only deniability or anonymity, respectively. The remainder (22.6%) want some combination of the three properties. When asked for an example of when they needed to use the properties, 98% of participants required non-repudiation at some point, and 82% of participants said it was very important to them. Whereas 60.94% of participants required deniability, only 23.18% mentioned that deniability was very important to them when needed.

- **Cryptographic deniability has not been considered by the courts when considering WhatsApp chat as evidence** We analyzed 228 US court cases where WhatsApp chats were part of the evidence. None of the cases presented an argument for cryptographic deniability. Even though some defendants claimed it was possible to forge messages, judges demanded evidence rather than accepting those claims at face value. We need court cases that present valid technical arguments for deniability in real-world instances to determine whether deniability will be legally accepted. Since we found no US court cases involving WhatApp that consider deniability, users are vulnerable to Harm 1 for apps that support deniability (*e.g.,* WhatsApp).

All the participants in the study were from the US and the legal analysis includes only US court cases. The preferences and expectations may differ substantially in other regions.

2 Background and Related Work

Systems provide different authentication properties.

(1) Non-Repudiation: A message has the sender's identity cryptographically bound to it so a third party has proof of who

sent it. A digital signature is a common method to achieve non-repudiation (*e.g.,* PGP [26], S/MIME [17]).

(2) Anonymity: The sender's identity is not bound to a message; the recipient has no evidence who sent it (*e.g.,* Tor [22]).

(3) Deniability: The recipient of a message can verify it came from the sender. However, the sender's identity is not cryptographically bound to the message, so the recipient cannot prove who sent the message to any third party.

In secure communication, the OTR protocol [2] first introduced cryptographic deniability using Deniable Authenticated Key Exchange (DAKE) [4]. Later Unger and Goldberg [23] improved DAKE to provide strong deniability for secure messaging (*i.e.,* IM apps). Other research formally analyzes deniability in Signal [24] and adds deniability to group messaging [20]. Besides secure messaging, research has explored deniability in other applications (*e.g.,* file systems [11], anonymous communication [12], document recommender system [25] and privacy-preserving data synthesis [1]). The lacuna in deniability research is understanding how users (and society at large) perceive deniability. Unawareness can lead to social engineering attacks (*e.g.,* see Section 5.2.1).

We are aware of only one other recent study besides ours that explores user understanding of deniability. Reitinger *et al.* [18] surveyed users to understand how they can be made more aware of deniability. The survey explored how different types of evidence affect deniability perception. The survey instructed participants to assume they were part of a jury in a court case where a hypothetical politician was accused of accepting bribes. The evidence for the accusation was a screenshot of the politician's messaging history.

Like ours, the study seeks to understand whether users accept the deniability provided by IM apps. Both studies show that participants do not accept deniability when presented with a screenshot of a conversation. We saw the same behavior even when we asked about specific contexts for a conversation (*e.g.,* close friend, untrusted acquaintance).

Our study differs in several ways from Reitinger *et al.* They used a courtroom setting to understand how users might perceive deniability while we focus on the issue in their personal lives. Our study considers different "contexts" that could significantly impact the acceptability of the claim. For instance, the relationship between the claimant and the recipient, screenshots of the chats versus showing messages directly in-app, *etc.* Also, our goal was to understand the users' current perception of deniability and what users expect from their Internet communication in daily life. To shed light on the legality of deniability, we studied actual US court cases. At a high level, our approach can be summarized in Figure 2.

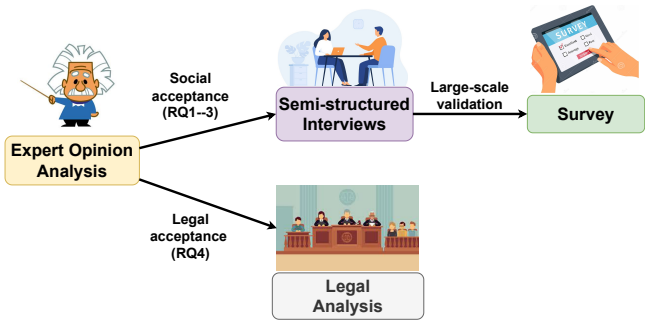


Figure 2: Overview of our approach.

3 Expert Opinion Analysis

To help formulate our research questions, we analyzed a thread [9] from `moderncrypto.org`—public forums for discussing modern cryptographic practice. The thread had 81 messages sent December 10–14, 2014, between 19 cryptography and usable security experts discussing the value of deniability in OTR-like protocols. Using participants' names and email addresses, we consulted their public web pages to determine their expertise.

3.1 Methodology

We used open coding to identify the topics discussed in the thread, followed by thematic analysis. Two researchers coded the discussion thread together and discussed and reconciled any discrepancies. Our intent was not to draw generalizable conclusions about the prevalence of specific issues. Instead, our analysis helped us determine the pros and cons of deniability. This understanding helped us formulate our research and design questions for our semi-structured interviews.

3.2 Results

Our analysis categorized expert opinions as (1) advocating deniability and (2) having reservations about its use and effectiveness. Both groups provided the reasoning for their positions. Experts in favor of deniability argued that:

E1: It mimics the expectations of in-person private conversation over a digital medium.

E2: It aligns with users' expectations because, in the past, unencrypted applications were deniable.

E3: It gives the sender a strong sense of security; the recipient cannot provide proof to any third party who sent the message. This property holds great significance for journalists and whistleblowers.

In contrast, experts not in favor of deniability argued that:

E4: Often, peers must prove to a third party that a conversation occurred between them. For instance, IM apps are used extensively in business and social life. Thus, non-deniable IM chats serve as proof of a business deal.

E5: In practice, it is challenging for ordinary users to forge messages in IM apps. Thus even if deniability were available to the users, they would be reticent to deny they sent a message because they doubt others will view a claim that the recipient forged a message as credible.

E6: Since users are largely unaware of deniability and its benefits, there is no social and legal acceptance. Lawyers who are unaware do not make strong arguments about the forgeability of messages (due to deniability) when courts consider chats as evidence.

Some experts suggested raising user awareness to achieve social (and legal) acceptance, such as providing an interface for users to forge messages easily. Others felt increasing cognitive load could negatively impact the user experience and lead to errors. Some experts believed in providing deniability without making users aware to minimize cognitive load, similar to current support for perfect forward secrecy. However, the effectiveness of deniability depends on the sender's awareness and the third party's acceptance. Therefore, users must be involved for deniability to benefit users. Teaching users about deniability is challenging. An expert reported: *"The way the OTR home page presents deniability is indeed confusing to users and can lead them to think they have some sort of extra protection in court when they don't."*

4 Semi-structured Interviews

The results of our expert analysis led to the development of our four research questions. We conducted online semi-structured interviews (N=12) to explore answers to RQ1–RQ3. Participants were from the United States. Each interview lasted between 40–60 minutes, and participants were compensated the equivalent of 13 USD per hour for their time.

4.1 Methodology

We designed our semi-structured interviews to contain the following three sets of questions:

RQ1—Social acceptance Some experts mentioned that deniability is essential to mimic in-person oral communication where a listener cannot prove to a third party what the speaker said. We asked questions to assess participants' trust in claims made through oral communication and messaging app chat. We asked questions from two perspectives: (1) participants as the recipient or listener of a message and (2) participants as third parties. For the first perspective, we asked participants whether they could prove to others that a message originated from a given sender. If so, how and why? For the second perspective, we asked participants if and why they trust a claim made by a recipient or listener.

We also asked participants to draw a diagram (while thinking aloud) showing all the entities involved when a message flows from a sender to a recipient. We used these diagrams to analyze whether participants' understanding of message flow correlates with their trust level in messaging apps.

RQ2—User understanding We showed participants the following OTR deniability definition and asked them to explain it and suggest use cases. *The messages you send do not have digital signatures that are checkable by a third party. Anyone can forge messages after a conversation to make them look like they came from you. However, during a conversation, your correspondent is assured the messages he sees are authentic and unmodified.* We used the definition because OTR was the first protocol to introduce deniability, OTR was the only messaging application with a deniability definition on its webpage, and an expert in the forum thread mentioned the definition could give users a false sense of security.

RQ3—Authentication property preferences We asked participants for their preferences regarding authentication properties (deniability, non-repudiation, and anonymity) for their Internet communication. To help ensure users understood the properties, we gave them an example use case for each property and encouraged them to ask questions.

We explained that deniability might be valuable to vulnerable groups (*e.g.,* whistle-blowers and journalists) and asked participants about their views on the importance of deniability for these groups. Also, we asked participants whether they wanted to use only one property or a combination (*e.g.,* non-repudiation for some messages and anonymity for others). Do they want these properties enabled system-wide or specific to just some apps?

Pilot study We conducted pilot interviews to (1) help us refine and gain experience with interview questions, (2) see how long an interview takes, and (3) improve the interview script and procedure, such as how to send them the OTR definition, how to explain deniability, etc.

Recruitment Participants were recruited using Prolific [3]. For ease of communication, we selected participants from the US who spoke English fluently. Also, we limited our study to participants who had used a messaging app (WhatsApp or Signal) and email service. To obtain diverse opinions, we tried to balance males (n=5, 42%) and females (n=7, 58%). Participant age ranges were fairly diverse: 21–30 (n=4), 31–40 (n=5), 41–50 (n=1), and 51–80 (n=2).

Data analysis We first transcribed the interviews using `otter.ai` [13]. One researcher read all the transcripts to fix transcription errors. Then, we analyzed the responses using inductive coding and content analysis [6, 10] by using the Quirkos tool [16]. To ensure inter-rater reliability, both researchers together coded all the interviews and discussed and reconciled all coding discrepancies. The complete codebook is available at `https://bitbucket.org/isrlauth/deniability/src/master/codebook.csv`. We conducted the interviews until we reached saturation.

4.2 Results

4.2.1 RQ1—Social acceptance

Participant as recipients When asked how they would prove to a third party that person A sent them a message, participants mentioned three alternatives: (1) show a screenshot, (2) forward the message, and (3) show the chat on the phone. Option 3 is preferred when the third party is untrustworthy, but otherwise, Option 1 had the most support (10/12).

We asked participants how their approach might change if the third party is a *court*. Eight agreed that chat messages in the app could be presented as evidence, two reported screenshots or video recordings of the messages are enough to be presented as evidence, and two mentioned the messages would only be circumstantial evidence. However, some felt courts must take additional steps to confirm the authenticity of the chat before submitting it as evidence, such as (1) subpoena WhatsApp company for a chat transcript, (2) getting a chat transcript from an ISP, and (3) forensic analysis of the chat.

Participants as the third party We presented participants with a scenario: *friend A* claims that *friend B* was talking badly about them (the participant). Would they believe *friend A*? No participant trusted the claim unconditionally. Six participants did not trust A because, ideally, A should have challenged B directly. The other participants said trust in A's claim would be based on the context of the claim, past experience, and A and B's motivations.

Next, we changed the scenario to assume *friend A* shows the chat on the phone (in person) to the participant. Nine participants completely trusted A's claim even if A has a history of lying and B is a close friend or family member (trustworthy). The high trust in the revised scenario is due to the participants' belief that (1) WhatsApp is secure due to robust encryption, so it is impossible to forge messages, and (2) A is not smart enough or has no motivation to put the significant technical effort required to forge the messages.

> **P6:** *"Even if friend A had a history of making stuff up, or not being the most trustworthy, she's showing me the texts on her phone what friend B said, then I'd be like, Okay this is legitimate."*

The interview results show that participants' expectations of messaging apps differ significantly from oral communication, contrary to claim *E1*. Participants have high trust in the integrity of the WhatsApp chat. Awareness of IM apps led participants to believe it was hard to forge messaging, contrary to claim *E2*. The results support claim *E5* that deniability will only be effective if messaging apps provide a way to allow users to forge messages easily.

Perception about Instant Messaging communication Participants have a different understanding of where messages flow from the sender to the receiver. They think messages can flow through (1) a centralized server, (2) a WhatsApp server, (3) a decentralized path, and (4) a direct connection.

Participants who believed there is a direct communication path between sender and receiver expect no one can read/modify their messages. However, some mentioned that government is an exception (see Figure 3). Overall, users' perception of WhatsApp as a secure E2EE app leads them to trust the messages more as a third party, which hinders the social acceptance of deniability.

Figure 3: P11's diagram for the flow of messages from a sender to a recipient on WhatsApp.

4.2.2 RQ2—User understanding

Only two participants correctly interpreted OTR's definition of deniability. The rest were confused or misunderstood it.

> **P2:** *"Well, I think this definition is saying two things that contradict..."*

> **P3:** *"Well, none of it makes sense"*

Two participants felt deniability prevents sender identity verification.

> **P9:** *"It looks like there's nothing to confirm in any way that the person you're talking to is a person you think you're talking to. To make it even worse, this can be altered..."*

4.2.3 RQ3—Authentication property preferences

When asked if they preferred deniability, non-repudiation, or anonymity, almost all participants (11/12) responded that they want non-repudiation because it provides deterrence against spam messages/misinformation and messages can be used as strong proof in court, *e.g.,* business deals over WhatsApp. Five participants believed anonymity was undesirable and had very strong views against it. Even after explaining that deniability could benefit vulnerable groups, participants preferred that non-repudiation should be the default property of all IM apps, and deniability could be optional. One participant stressed that they would feel insecure if deniability was the default, and another mentioned that:

> **P12:** *"I do like that there are ways to communicate online anonymously and disputably. I think that it can be a really good thing for some people, but I feel like a lot of where this country is, it's partially because of the lack of any control and any fact-checking online. So I would go with indisputable to help combat this problem..."*

Combination (and nuanced features) of the authentication properties Participants exhibited a wide range of preferences when asked about using combinations of properties.

1. *Person-based:* The property depends on the contact. For instance, when Bob sends messages to Alice, they would be non-repudiable, but when the recipient is Carol, messages would be deniable.

2. *Message-based:* Irrespective of the recipient, some select messages could be deniable and some not. For instance, if Bob is messaging Alice, Bob can selectively decide to send some messages that are deniable and some that are not. Importantly, participants who suggested this property wanted non-repudiation as the default property default but other properties (*e.g.,* deniable, anonymity) as optional.

3. *App-based:* the property depends on the app being used (*e.g.,* Gmail communication could be deniable, and WhatsApp messages could be non-repudiable).

4. *Time-based:* For a specified time (*e.g.,* one hour), messages sent by the participant would hold a particular authentication property. *E.g.,* the messages are non-repudiable for one hour, and then they will become deniable.

5. *Duration-based:* For a specified duration, any message from the participant would have a specified property. For example, for the next two hours, all messages from a user will be non-repudiable (and they will always remain non-repudiable). But after two hours, the messages will be sent with a different authentication property.

5 Survey

We conducted a survey (**n=664**) of users recruited via the Prolific platform. Our goal was to quantify how users perceive messaging app communication when given different contexts for the conversation and what authentication properties they expect from their Internet communication applications.

5.1 Methodology

A total of 931 participants attempted the survey and 731 finished it. We discarded 67 responses that failed the concentration check, leaving 664 responses in the final analysis.

To personalize the survey questions, we asked the participants to enter the names of people belonging to different categories based on trust (a close friend, an untrustworthy person *etc.*). Later during the survey, we asked the participants to correctly categorize the name they initially provided. If they failed, we discarded their responses due to a lack of concentration. Because of our strict concentration check, we paid all 731 participants (including those who failed the concentration check). We compensated participants at the rate of $13.46 per hour. The medium time to complete the survey was 13 minutes and 40 seconds, and we paid each participant $3.05.

Metric	Percent	Metric	Percent
Gender		**Ethnicity**	
Male	46	White	57.3
Female	45	Asian	12.2
Age		Black	10.4
18-29 years	30.3	Mixed	5.7
30-39 years	34	**Employment Status**	
40-49 years	17.2	Full-time	42.7
50-59 years	11.2	Part-time	10.9
60+ years	7	Unemployed	7.1
Student status		Unpaid work	4.9
Student	20.7		
Non Student	54.6		

Table 1: Survey participant demographics. Percentages may not add to 100% because we do not include "Other" or "Prefer not to answer" percentages for brevity.

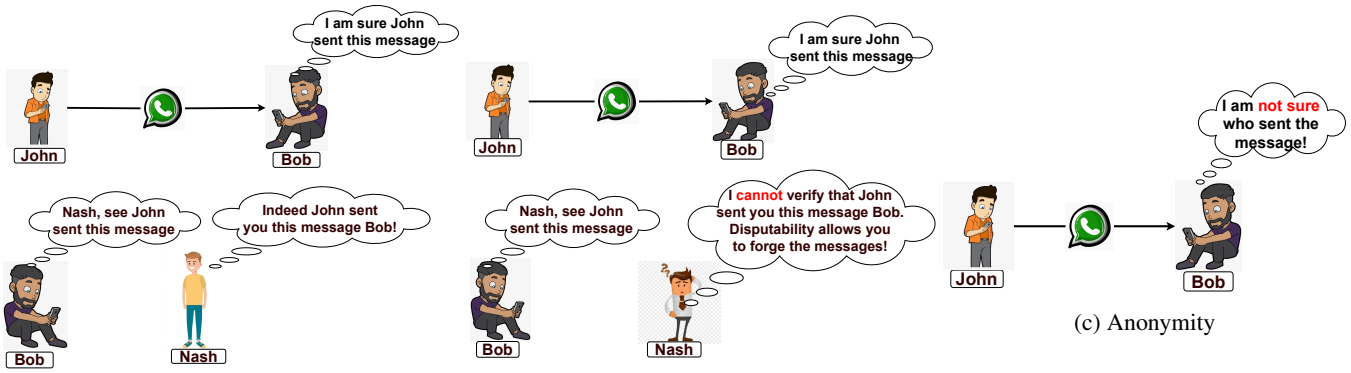
Table 1 summarizes our participants' demographics. We had a gender-balanced distribution; our participants' age and ethnicity distribution was similar to that of the US population.

We conducted the pilot study in two phases to improve our survey questions: Phase 1 (n=13)—friends, family, and co-workers and phase 2 (n=29)—IRB-approved survey on Prolific. We asked participants to think aloud while answering questions to ensure they interpreted our questions correctly. Participants correctly interpreted our questions. They understood our description of non-repudiation and deniability definitions presented in the pilot survey.

After the pilot survey, we revised our survey questions to conduct the survey with **n=664** participants. The complete survey is available at https://bitbucket.org/isrlauth/deniability/src/master/survey.pdf.

RQ1—Social acceptance To measure participants' perceptions of deniability in messaging applications, we devised three distinct scenarios representing different roles and formulated questions for each scenario. Additionally, we inquired about participants' beliefs regarding the forgeability of messages to explore any potential correlation between their perception of deniability and their beliefs about message forgeability.

Participant as a third party: Alice (a recipient of a message or listener) claims to the participant that Bob (sender or speaker) was talking badly about the participant. We asked participants to provide trust scores **(1= no trust, 10 = complete trust)** for a different combination of the following factors: (1) Alice's relationships with the participant: a trustworthy person (*e.g.,* family member, close friend), an acquaintance, or an untrustworthy person, (2) Bob's relationships



(a) Non-repudiation (Indisputable)

(b) Deniability (Disputable)

(c) Anonymity

Figure 4: Different Authentication Properties.

with the participant (same as aforementioned), and (3) how Alice makes their claim: orally, by showing a screenshot of the chat, or by showing the chat in the messaging application.

Participant as the receiver: First, we asked participants a scenario-based question to identify whether they, as a receiver, consider a messaging application chat as evidence. The scenario asked participants if they could use a message from their landlord as evidence. Second, we verified if deniability can be exploited to deceive participants into believing something they never said. We presented participants with a scenario where Mallory claims that the participant owes Mallory $50 from dinner, but the participant does not remember. We asked participants to provide trust scores (1= no trust, 10 = complete trust) for a different combination of the following factors: (1) Mallory's relationships with the participant and (2) how Mallory makes their claim (orally, showing a screenshot, or in-app chat).

Participant as the sender: We asked participants how the receiver could prove to a third party that the participant had sent them a message.

Participant beliefs regarding message forgery: We asked participants about their beliefs on how hard it is to forge a screenshot and a chat in a messaging app.

RQ2—User understanding We presented participants with OTR's deniability definition verbatim and asked them to choose among several statements about deniability, including incorrect interpretations from interview participants.

RQ3—Authentication property preferences We asked participants direct questions about authentication properties they expect for their Internet communication. To make it easier for participants to contrast between non-repudiation and deniability, we used the terms "indisputable" and "disputable" with our own explanations. We also displayed diagrams for each of these properties to help them understand (see Figure 4). In our pilot surveys, all participants correctly understood our explanations.

To indirectly infer the need for these properties in participants' lives, we asked them to describe instances where they needed non-repudiation and deniability. If they ever needed these properties, we asked them to rate the importance of achieving the corresponding property in each scenario.

5.2 Results

5.2.1 RQ1—Social acceptance

Participants as third party As previously explained, we provided participants with a scenario where Alice tells them that Bob was talking badly about the participant behind their back. We analyzed the change in the participants' trust scores across three factors:

- Participant's relationship with Alice (*i.e.,* recipient/claimant): Alice is (1) a trustworthy person 'T', *e.g.,* a close friend, (2) an acquaintance 'A', or (3) an untrustworthy person 'U', *e.g.,* a person with a bad history.
- Participant's relationship with Bob (*i.e.,* sender about whom Alice makes a claim). Same as described above.
- Medium through which Alice claims to the participant (*i.e.,* third party): orally, by showing a screenshot of the chat, or by showing the chat in the app. These are represented as 'O', 'S', and 'I', respectively.

Figure 5 represents the participants' mean trust score on the Y axis and different combinations of factors on the X axis. Each point on the X-axis is a three-tuple value <Participant's relationship with Alice, Participant's relationship with Bob, medium through which Alice conveys the claim to the participant>. For example, tuple `<T-T-O>` represents that Alice is trustworthy (T), Bob is also trustworthy (T), and Alice orally (O) tells the participant that Bob spoke badly about the participant. In this case, the mean trust score on the claim by our participants (as a third party) was 7.31.

Medium of communication impacts deniability When both Alice and Bob are trustworthy, how Alice makes their claim to

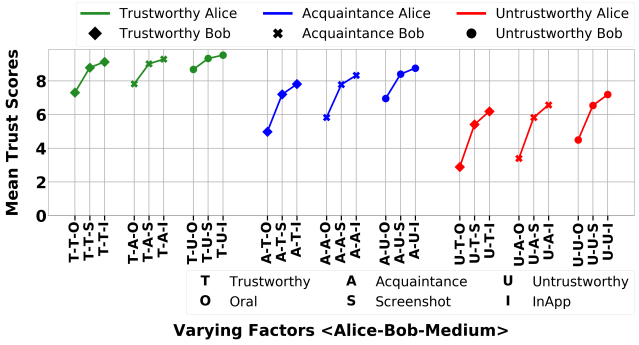


Figure 5: Third-party trust in a recipient's claim for different combinations of human relationships and mediums of communication.

the participant impacts the trust score. Overall, when a claim is made by showing a screenshot, the mean score increases to 8.79, and when Alice shows the messages in the app, it reaches 9.13. The trust score is highest when a claim is made by showing a chat in the app, and this pattern also holds for all other relationships. Moreover, there was a significant trust change from oral to screenshot (1.78) and a relatively smaller trust change from screenshot to in-app (0.50), showing that people's trust in screenshots is close to in-app.

The results show that third parties trust screenshots and in-app chats significantly more than oral claims, raising questions about whether those who want to use deniability in messaging apps can do so successfully.

However, 4.2% (28) participants trusted oral claims more than screenshots or in-app. It appears these participants doubted a claimant if they make an inordinate effort to prove their claim, as mentioned by a participant in our interviews. On the other hand, there was no difference in trust between oral, screenshot, and in-app claims for 9.2% (61) participants. Future in-depth studies are needed to explore the mental model of these participants and what led them to accept deniability on WhatsApp (un)intentionally.

Human relationships impact deniability For tuple <T-U-O>, the trust score is 8.68, whereas, for <T-T-O>, it is 7.31. This difference shows that if a trustworthy person makes an oral claim about an untrustworthy person, participants find the claim more convincing than a claim about another trustworthy person. A similar pattern also holds for other mediums, suggesting that human relationships impact a third party's acceptability of the claim.

The interplay of human relationships and medium of communication The change in trust score between different relationships varies depending on the underlying medium. For example, consider all the cases of trustworthy Alice (the three green lines) and untrustworthy Alice (the three red lines). We first computed the difference between when trustworthy Alice *orally* claims to the participant and when untrustworthy Alice *orally* makes the same. To ignore the effect of Bob's relationship, we computed the trust score difference for the same Bob's relationship (<T-**T**-O> − <U-**T**-O>, <T-**A**-O> − <U-**A**-O>, <T-**U**-O> − <U-**U**-O>.) We calculated the mean of the differences, giving us the mean trust score difference (4.34) when trustworthy and untrustworthy Alice make the *oral* claim.

Next, following a similar procedure, we computed the difference between when (1) trustworthy Alice claims showing an *in-app chat* to the participant and (2) untrustworthy Alice makes the same claim showing an *in-app chat*. This difference is 2.67, which indicates that change in human relationships has a relatively lesser impact when the claim is made in-app than orally. A paired sample t-test shows a significant difference between a change of trust based on a relationship orally (M=4.34; SD=2.69) and in-app (M = 2.67; SD=2.65); [t(660)=16.881, p<0.001].

We previously established that the way the claim is made to a third party significantly impacts the acceptability of the claim. But Alice's relationship also plays a role in the acceptability of the claim. The highest increase in trust between oral and in-app occurs when an untrustworthy person claims about a trustworthy person (<U-T-O> and <U-T-I>). For the oral claim, the trust score is 2.89, and it changed to 6.19 when the claim was made by showing in-app messages. The drastic increase in trust scores on an untrustworthy person makes users vulnerable to social engineering attacks, as untrustworthy people are more likely to launch the attacks.

Through our multi-perspective scenarios, we observed that different factors, like human relationships and how the claim is made to a third party, affect the acceptability of the claim and, in turn, the deniability. We found that people's trust scores on claims made in-app are significantly higher than those made orally.

Statistical analysis To determine whether the change in trust scores is statistically significant, we performed repeated measures two-way ANOVA test. It allows us to see an interaction between two independent variables, the medium and the third party's relationship with the claimant. Due to the violation of the assumption of sphericity, tested by Mauchly's test ($\chi^2(9) = 1173.114, p < .001$), we used Greenhouse-Geisser correction. We did not check the normality assumption because with large enough sample sizes (> 30 or 40), the violation of the normality assumption should not cause major problems [7]. A two-way ANOVA revealed that there was a statistically significant interaction between the effects of the medium and the relationship with the claimant (F(2.207, 1454.157) = 166.236, p<.001).

To determine whether there was a simple main effect for the medium or relationship, we ran two repeated measures one-way ANOVA on each relationship and medium. As shown in Table 2, there was a statistically significant difference within the mediums and also within the relationships. Next, we examined pairwise comparison as shown in Table 3. We found that there was a statistically significant difference in

Table 2: Repeated measures ANOVA on each relationship and medium – Participants as a third party.

	Mean trust score						
Relation	I	S	O	df	F	Sig.	η^2
T	9.33	9.06	7.94	1.31, 865.015	277.76	<0.001	0.297
A	8.31	7.80	5.93	1.43, 950.96	665.13	<0.001	0.502
U	6.65	5.93	3.60	1.47, 966.750	755.879	<0.001	0.534
Medium	T	A	U	df	F	Sig.	η^2
I	9.33	8.31	6.65	1.52, 1005.017	515.81	<0.001	0.439
S	9.06	7.80	5.93	1.58,1041.75	683.56	<0.001	0.509
O	7.94	5.93	3.60	1.79, 1181.82	1164.53	<0.001	0.638

T = Trustworthy, A = Acquaintance, U = Untrustworthy
I = InApp, S = Screenshot, O = Oral

Table 3: Pairwise comparisons from one-way ANOVA on medium and relationship.

Comparison	Mean Difference	Std. Error	Adj. Sig	Lower Bound	Upper Bound
InApp-Oral	2.276	0.073	<0.001	2.101	2.452
Screenshot-Oral	1.775	0.063	<0.001	1.624	1.925
InApp-Screenshot	0.502	0.039	< 0.001	0.408	0.596
Trustworthy-Untrustworthy	3.379	0.093	<0.001	3.157	3.601
Acquaintance-Untrustworthy	1.947	0.074	<0.001	1.77	2.123
Trustworthy-Acquaintance	1.432	0.062	<0.001	1.283	1.582

the trust score between all pairs of mediums and all pairs of relationships: (1) in-app > screenshot > oral, and (2) trustworthy > acquaintance > untrustworthy. Note that we used the Bonferroni correction for multiple tests.

Participants as receiver To understand the participants' perception as recipients, we presented a scenario where, as recipients, they could believe that WhatsApp chat is sufficient evidence to prove that a sender sent them a message. 32.7% agreed that WhatsApp chat is sufficient to use as evidence.

The interviews revealed an interesting case where an attacker deceives a person by claiming the person sent the attacker a message. To determine the likelihood of this threat, we asked a scenario-based question where Mallory claims to the participant that the participant owes them $50. Mallory makes this claim orally, showing a screenshot (where the participant acknowledged owing $50) and an in-app message (acknowledging the same).

Figure 6 shows that trust in a claimant and how they make a claim significantly impacts how the receiver perceives the claim. Following the same notation used earlier, the mean trust score of the participants always increases when an oral claim is accompanied by a chat shown in the app.

Statistical analysis To determine whether the change in trust scores is statistically significant, we performed repeated measures two-way ANOVA Test, which allows us to see the interaction between two independent variables (Channel, Relationship with the claimant). Due to the violation of the assumption of sphericity, tested by Mauchly's test ($\chi^2(9) = 1064.775, p < .001$), we used Greenhouse-Geisser

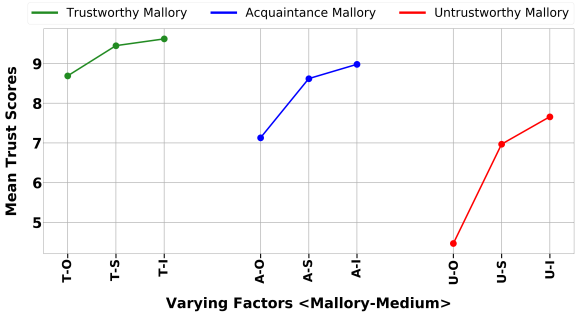


Figure 6: Users' trust in deceiving claims where an attacker convinces users that they made some statements in the past. The graph shows the impact of different human relationships and mediums of communication on trust scores.

Table 4: Repeated measures ANOVA on relationships and mediums—Trust score for false claims with chat evidence of a past statement.

	Mean trust score						
Relation	I	S	O	df	F	Sig.	η^2
T	9.63	9.46	8.70	1.23, 853.304	153.102	<0.001	0.189
A	9.00	8.63	7.13	1.37, 896.69	422.36	<0.001	0.391
U	7.67	6.97	4.48	1.46, 961.52	690.212	<0.001	0.512
Medium	T	A	U	df	F	Sig.	η^2
I	9.63	9.00	7.68	1.42, 932.83	290.62	<0.001	0.306
S	9.46	8.63	6.97	1.46, 957.76	414.31	<0.001	0.386
O	8.70	7.14	4.47	1.74, 1150.18	915.83	<0.001	0.581

T = Trustworthy, A = Acquaintance, U = Untrustworthy
I = InApp, S = Screenshot, O = Oral

correction. A two-way ANOVA revealed that there was a statistically significant interaction between the effects of the channel and the relationship with the claimant ($F(2.388, 1569.017) = 247.117$, $p < .001$).

To determine whether there was a simple main effect for the relationship, we ran repeated measures one-way ANOVA on each channel. As shown in Table 4, there was a statistically significant effect of the relationship on trust for each channel.

By examining the pairwise comparisons using one-way ANOVA, see Table 5, we found a statistically significant difference in the trust score between all pairs of mediums and all pairs of relationships: (1) in-app > screenshot > oral, and

Table 5: One-way ANOVA pairwise comparisons of channels and relationships—Trust score for false claims with chat evidence of a past statement.

Comparison	Mean Difference	Std. Error	Adj. Sig	Lower Bound	Upper Bound
InApp-Oral	1.996	0.072	<0.001	1.824	2.168
Screenshot-Oral	1.587	0.062	<0.001	1.438	1.735
InApp-Screenshot	0.409	0.036	< 0.001	0.322	0.496
Trustworthy-Untrustworthy	2.890	0.098	<0.001	2.655	3.124
Acquaintance-Untrustworthy	1.875	0.075	<0.001	1.694	2.056
Trustworthy-Acquaintance	1.014	0.059	<0.001	0.872	1.156

(2) trustworthy > acquaintance > untrustworthy. Note that we used the Bonferroni correction for multiple tests.

Participants as sender Our survey asked a scenario-based question inquiring if participants, as senders, think that they can use WhatsApp messages as evidence that they sent to a particular receiver. 63.25% of all participants agreed that they could use WhatsApp chat as proof of the conversation.

Participants' belief on forging the message We asked participants to indicate how easy they believe it is to forge (1) a screenshot of a chat and (2) a message within the application. Figure 7 shows that most participants believe forging a message within the app is much harder than forging a screenshot.

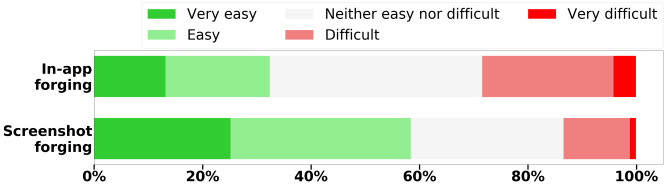


Figure 7: Forging the screenshot vs. chat in the app itself.

To measure the correlation between a change in trust score and participants' belief about the difficulty of forging an in-app message (or a screenshot), we computed the Pearson correlation coefficient (r). We found the trust score change between oral and in-app and participant's belief about forging messages to be weakly correlated *i.e.,* $r(664) = 0.114, p = 0.003$. Weak correlation shows that even if users believe a screenshot and chat are easy to forge, they trust them significantly more than oral claims. Increased trust could be due to users assuming they are unlikely to suffer a forgery attack.

Summary: Participants as a third party tended to have significantly more trust (2.28) in in-app chats than oral claims, raising questions about the feasibility of successfully utilizing deniability in messaging apps. As a recipient, 32.7% believed they could use a received message as proof. Whereas as a sender, 63.25% believed they could use chat as evidence of a conversation.

5.2.2 RQ2—User understanding

Because the interview participants completely misunderstood OTR's deniability definition, we sought to quantify these misunderstandings through our survey. Therefore, we gave the survey participants the OTR definition verbatim and provided some statements regarding message authenticity. We asked the participants to indicate which statements were true based on the definition (see Figure 8). There was only one correct statement; the rest were common misunderstandings we observed during our interviews. Figure 8 shows that most participants marked all the statements as true; only 0.6% of the participants selected *only* the correct statement (*i.e.,* second statement).

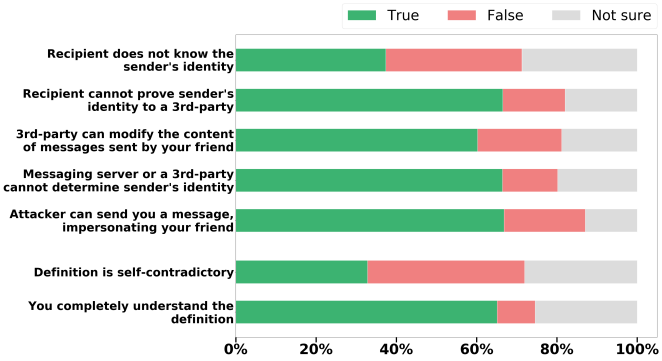


Figure 8: Users' understanding about deniability property.

Interestingly, around 70% of the participants thought they understood the definition but did not. This gap could lead to a false sense of security or a bad reputation for deniability. For example, suppose a user thinks they understood the definition to imply a messaging app cannot track the sender, so they falsely believe they cannot be held responsible for what they say in the app. In addition, if a user thinks an attacker can forge a message from their friend, they might refuse to use apps supporting deniability.

5.2.3 RQ3—Authentication property preferences

Figure 9 shows that more than 60% of users *solely* desire non-repudiation, 12.7% desire deniability, and 4.5% desire anonymity. For participants who desire some combination of properties (22.6%), we asked them for their preferred approach from the choices identified during our interviews (See Section 4.2). Figure 10 shows the most popular approaches are application-based (42%), message-based (32%) and person-based (23%). Only 2% prefer time-based.

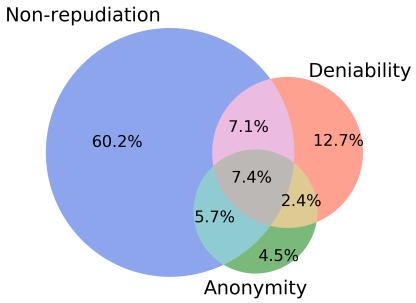


Figure 9: Percentage of users preferring a different combination of authentication properties.

The interviews revealed that some users prefer different authentication properties for different applications. Therefore, we asked participants to select their most preferred property

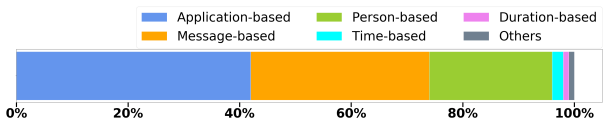


Figure 10: Users' preferred approach to combining authentication properties.

for three application categories. The majority (> 60%) selected non-repudiation for each category (see Figure 11).

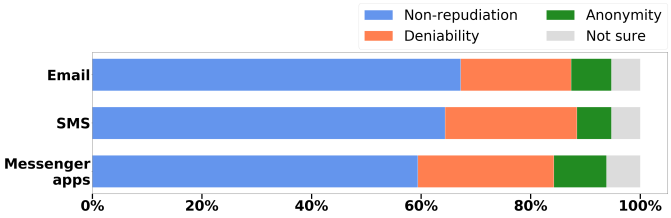


Figure 11: Users' expectations regarding authentication properties for different messaging apps.

Our analysis further revealed that some participants want a single property across all three application categories: 45.9% non-repudiation; 8.1% deniability, 1.7% anonymity, and just 3% are unsure what property they want. The remaining participants (41.3%) desire different properties for different application categories, but there is no consensus on the combination of properties they want. Each of these combinations was preferred by less than 3.2% of participants.

To measure the practical need for deniability and non-repudiation, we asked participants to describe a real-life scenario for each property and how important it was for them to have it; 39.1% of participants had never needed deniability, whereas only 2.2% had never needed non-repudiation. Figure 12 shows that 82% of the respondents reported that non-repudiation was very important for their scenario. In contrast, only 23.2% reported that deniability was very important for the scenario. These results show that non-repudiation is very important to more people than deniability.

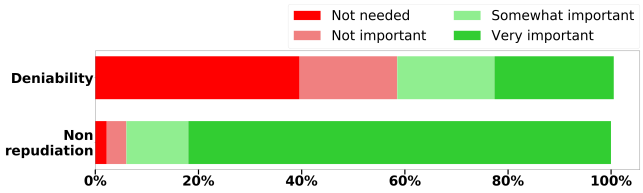


Figure 12: Importance of non-repudiation and deniability for the participants.

However, 18.9% of participants mentioned that deniability was somewhat important (if not very important) for their scenarios. In total, 42% of participants consider deniability at least somewhat valuable. Interestingly, our findings show more participants *i.e.,* 42% *needed* deniability than those who *preferred* it *i.e.,* 29.6% (see Figure 9). These results align with an expert's opinion from the public forum [9] that some users might not know what they need.

6 Legal Analysis—RQ4

To understand the legal acceptance of deniability, we analyzed US court cases where WhatsApp chat was considered evidence. At the recommendation of several law professors, we hired a senior law student from the BYU Law School who had already completed suitable courses for our study *viz. evidence* and *contract*. She helped us retrieve relevant court cases and interpret them accurately.

6.1 Methodology

Data gathering We used Westlaw [19], an online legal research service and proprietary database for lawyers and legal professionals, to retrieve candidate court cases (judicial opinions). Our goal was to study cases where IM apps were brought up as evidence. We focused on WhatsApp because of its popularity in the US [5]. When we directly searched for the term "WhatsApp" on the Westlaw search portal, in most of the results, WhatsApp was not taken up as evidence. Instead, it was just mentioned for other reasons *e.g.,* WhatsApp was used to notify the parties involved in a lawsuit. To exclude irrelevant cases, we created search queries that return the cases where WhatsApp chat was considered admissible evidence and where it was rejected as evidence. To ensure that we obtained almost all cases, we consulted our legal expert and Westlaw support center when constructing the queries. Please refer to Appendix A.1 for our Westlaw search queries. We retrieved a total *228* unique court cases for analysis.

Data analysis We performed a qualitative and quantitative analysis of the 228 court cases. First, two researchers and the legal expert independently studied all the cases and classified them into one of the following four categories based on the importance of chat as evidence. Then, they discussed and resolved any conflicts to reach a consensus.

- Major evidence: WhatsApp chat that could make or break one side's case if it was the only evidence presented.
- Minor evidence: Circumstantial evidence that cannot lead to the win (or loss) of the parties involved. This evidence assists in the corroboration of the story but could not result in a decision for either side on its own.
- Rejected evidence: WhatsApp was brought up as evidence but rejected by a judge because of potential forgeability and deniability claims.

- N/A: WhatsApp is mentioned for purposes other than evidence, such as notifying the parties in a lawsuit.

Next, researchers performed a qualitative analysis of cases where WhatsApp chat was considered major or minor evidence. Two researchers together read all these cases and made a broad list of topics around WhatsApp as evidence. They finalized the list into themes to understand why WhatsApp is considered significant evidence even after it offers deniability property. Notably, our goal in analyzing these cases was not to draw generalizable conclusions about the prevalence of specific issues but to identify the reasons for considering WhatsApp as the evidence.

6.2 Results

Total cases	Evidence Categories			
	Major	Minor	Rejected	N/A
228	79 (34.6%)	76 (33.6%)	0	73 (31.8%)

Table 6: Evidence categories for WhatsApp chat court cases.

Table 6 provides our classification of the 228 cases into evidence categories of major, minor, and N/A (not mentioned as evidence). From our qualitative analysis, WhatsApp chat was never rejected as evidence because of deniability (or the possibility of a forged chat), but was considered major evidence in many cases. In one of the cases, the court concluded that:

> Hulsh v. Hulsh: *"Jeremy's central piece of evidence of Svarinsky's sexually inappropriate behavior is a WhatsApp message thread between Viera and Svarinsky..."*

In another case, the recipient's testimony and the messages were enough to convict a person of robbery.

> United States v. Rivas Nunez: *"Given Castillo Vallejo's testimony, the call log, and the contents of the WhatsApp messages, there was sufficient proof to enable a jury to determine by a preponderance of the evidence that Rivas Nunez was involved in a conspiracy to rob AT&T with Castillo Vallejo and Rodriguez Nunez."*

Forgery claims ignored Surprisingly, even when a sender claimed a WhatsApp chat was untrustworthy, it was considered primary evidence. In one criminal case, the defendant argued, "many of the messages are incomplete and cannot be authenticated, and WhatsApp messages generally are unreliable due to hacking vulnerabilities." Moreover, due to unauthenticated messages, they filed a 'Motion in Limine' to *not* present the WhatsApp chat as evidence in front of the jury. The court rejected the request; verifying the authenticity of the chats was left to the jury.

> United States v. Ojimba: *"Defendant's objection that the text messages, in this case, are unreliable is made without any persuasive evidence and is thus overruled... The court explained that Mr. Ojimba could attack the reliability of the messages at trial, but that reliability was ultimately a matter for the jury."*

In these cases, the lack of the defendant's (sender's) awareness of the deniability property in WhatsApp makes it harder for them to make a persuasive argument in court regarding the non-authenticity of messages in the app (*i.e.,* the recipient could have forged the messages). In the case of a jury trial, the lack of social acceptance of deniability among the jurors may influence their verdicts.

Prior history impacts deniability In some cases, a person's history (including job, character, *etc.*) impacts deniability. In one case, the judge mentioned that unrelated acts/history can be used for proving "motive, opportunity, intent, preparation, plan, knowledge, identity, absence of mistake, or lack of accident." Thus, using a person's past makes the *utility* of deniability harder in legal proceedings, even if it is socially acceptable. For instance, if a journalist denies they sent a message describing a scandal, the court may doubt the claim if the journalist has a history of exposing scandals.

Interestingly, the court even allowed using WhatsApp messages from the past to set up intent to convict people for different cases. The following quote shows the judge's response when the defendant asked to remove WhatsApp chat as evidence as it was unrelated to the case in consideration.

> United States v. Ramirez-Frechel: *"Because the messages showed William was engaged in the business of dealing in firearms with the purpose of making a livelihood and profit, the district judge did not abuse his discretion in admitting the WhatsApp messages that happened in the month after March 23rd as intrinsic evidence of the charged gun-dealing crime."*

Screenshots as evidence In some cases, even a screenshot was acceptable as long as the chain of custody confirms the screenshot was taken directly from WhatsApp.

> United States v. Avenatti: *"Ms. Clifford provided limited consent for the Government to screenshot in her presence certain portions of [her] WhatsApp text messages with the defendant, and an Investigative Analyst with the United States Attorney's Office did so. The Government's understanding is that the export was generated automatically using an electronic feature that compiles the entirety of a WhatsApp conversation into a printable and shareable electronic file."*

None of the analyzed cases employed digital signatures, and courts did not rely on them to establish message authenticity. Malicious individuals could exploit this potential loophole to fabricate false evidence through systems that provide deniability. In high-value criminal cases, defendants could call upon expert witnesses to testify and address deniability properties. However, in more minor civil cases, defendants or their

legal representation may lack the knowledge or resources to enlist such experts.

Deleted messages as evidence In some cases involving WhatsApp, the recipient showed screenshots of messages on their phone while the sender deleted them and claimed the screenshots were forged. Forensic analysts retrieved the messages on the sender's phone as evidence.

> United States v. Ozbirn: *"The analyst extracted messages Appellant sent to and received from Febes and Jodie via WhatsApp...The WhatsApp messages taken from Appellant's phone are nearly identical to the screenshots of Febes's and Jodie's phones..."*

Encryption implies criminal wrongdoing In one case, using WhatsApp raised suspicion because it supports encryption. A search warrant was requested for further investigation.

> United States v. Ciuca: *"The agent's affidavit to support the search warrant of the subject email account describes the WhatsApp chats found on Bitere's phone. The agent indicates that WhatsApp is frequently used by individuals engaging in crime due to its strong encryption."*

7 Limitations

Our research has the following limitations:

Demographics Although the demographic attributes of the participant group are close to the US average, Prolific participants do not reflect the general population. Also, all of our participants are from the US and may have different perspectives about deniability compared to other regions.

Human bias Our study may be susceptible to participants' bias as our scenarios were abstract, and participants were asked to imagine themselves in situations they may not have encountered. Despite these limitations, presenting multiple scenarios to participants allowed us to explore situations that might not currently happen but are similar to situations that could happen in the future.

Priming bias The findings on authentication preferences (RQ3) may be influenced by priming bias. Early in the interview, participants considered scenarios where they had to verify the sender's identity rather than conceal their identity. This may have biased later responses regarding their preference for authentication properties. Although we presented an example illustrating the usefulness of deniability before asking for their preference, more than a single example may have been needed to fully convey the benefits of anonymity or deniability to the participants.

Misinterpretations In our study, participants may have misinterpreted our deniability and non-repudiation definitions. Since our interviews revealed that participants did not understand OTR's deniability definition, we created definitions and showed them with an illustrative diagram during the pilot study (see Figure 4). The pilot showed that participants understood them, but a large-scale study would help to confirm the definitions are understandable.

Legal analysis We hired a senior student from the BYU Law School as an expert to help categorize court cases based on the WhatsApp weight of evidence. A professional, experienced lawyer's opinion might differ from our expert's opinion on the weight of evidence category. To minimize this limitation, we were conservative in our analysis. When in doubt, we placed the evidence in a lower category. Moreover, since our study had participants and legal cases from the US, our inferences and conclusions are not generalizable.

8 Discussion

Deniability is ineffective Although we did not study vulnerable populations (*e.g.,* whistleblowers), our results are highly relevant to them since deniability works only if it is accepted by society and the courts. Participants in social scenarios trusted in-app evidence significantly more than oral claims, implying chat conversations may be far less deniable. In legal scenarios, judges rely on WhatsApp chats as evidence and show no inclination to support deniability. Also, jurors may reflect the bias that they trust chats more than oral claims. The lack of acceptance in social and legal scenarios by default makes deniability ineffective for people who may need it.

Deniability is hard to achieve in legal cases Even if deniability is socially accepted, it appears much harder to achieve legal acceptance. A sender may deny sending a message in court, but it may still be on their phone as evidence. If they delete messages and deny sending them, a forensic analyst might still retrieve them (see Section 6.2). Other evidence outside the user's control can limit deniability, such as an ISP furnishing proof that Alice sent a message to Bob on a given date and time. Currently, users lack reliable information on what they should do if they want to deny sending a message. Simply deleting a message after sending it is not enough.

Users are vulnerable to social engineering when unaware of deniability Our results show a lack of social and legal acceptance of deniability for WhatsApp, even though it supports deniability. As a result, users are vulnerable to social engineering attacks (Harm 1), as shown in Figure 1. An attacker can forge a message to appear as if they received it. Third parties will trust it socially and legally. In general, a trustworthy person has more potential to deceive a user than someone untrustworthy. However, our results show that the largest increase in trust scores comparing oral to in-app occurred when an untrustworthy person made a claim. The more trust increases for in-app claims than oral claims, the more vulnerable users are to social engineering attacks from untrustworthy persons who exploit the deniability property.

Deniability has a bad reputation Participants believed bad actors would use an app to spread fake news and misinforma-

tion if it supports deniability. The misuse of deniability could overshadow its advantages. Thus, an area for future research is minimizing the risks associated with deniability.

Even a screenshot is not deniable Trust scores increased three times more from oral to screenshot than from screenshot to in-app. Users trust screenshots over oral claims despite believing they are easy to forge. Future research could explore why. Users treat screenshots like written content, which they trust more than oral claims—understanding why may lead us closer to the social acceptance of deniability in-app.

Takeaways for different stakeholders Application developers may need to revisit default behavior. Defaulting to deniability may be unsuitable and cause adverse effects due to a lack of user understanding. It prevents holding message senders accountable. Over 80% of the participants wanted non-repudiation, and 98% gave examples of needing it in practice. But, despite the majority preference, some users desire and require deniability. Future research can explore the coexistence of authentication properties in apps for diverse needs.

More user-friendly definitions of deniability are needed. Currently, OTR's definition leads to a false sense of security and a loss of app reputation.

Only 2% of participants preferred time-based deniability, challenging the suggestion to publish email providers' DKIM secret keys at regular intervals [8,21]. More study is needed to test different combinations or properties and their use cases.

We need to raise awareness with legal stakeholders that deniability allows bad actors to forge messages and use them in court since courts accept WhatsApp chat as evidence.

Combinations of authentication properties Even though participants preferred different authentication properties, there are challenges to providing alternatives. These include increased cognitive load, conflicting preferences between senders and receivers, attacks coercing non-experts to make choices that harm them, and increased complexity for application developers.

Future work Future studies can include participants from different countries and cultures. Since legal systems vary between countries, future studies should involve legal experts from the respective countries. Court cases can be searched to see if deniability has ever been supported in other jurisdictions. If results vary across countries, it leaves people vulnerable to different attacks described in Figure 1 than the risks we identified in the US.

Our results open up new research avenues for deniability, such as (1) raising awareness, (2) identifying and communicating positive use cases, and (3) increasing social acceptance to benefit vulnerable populations. Since vulnerable populations are most likely to benefit from deniability, a future study could focus on their perceptions of deniability.

Ethics We received IRB approval for the interview and survey studies. We asked survey participants for a pair of names from each relationship category to personalize the questions and encourage consistent responses across different questions. To minimize harm, we asked participants to enter only nicknames and permanently deleted names after the survey. We collected no personally identifiable information during the study. The IRB determined our analysis of expert opinions on a public forum did not require their approval.

9 Conclusion

We conducted a multi-perspective study of deniability, including an analysis of expert opinion, user interviews and surveys, and an analysis of court cases. The results show that deniability is not socially accepted in the US. It still needs to be determined whether deniability can be legally accepted since we found no US court cases where the defense made a case for deniability. The result is that deniability is ineffective for the people who may need it.

The potential for forensic evidence on a user's phone or ISP raises questions about whether cryptographic deniability will actually lead to the complete deniability of messages in legal settings. This is an interesting direction to explore.

Our survey shows that most users prefer non-repudiation over deniability, but some users think both are important. This raises the question of whether both can be supported to fill different user needs.

Our research illuminates the need for human-centered research in deniability supporting user choice in authentication properties.

Acknowledgments

The authors thank the reviewers and our shepherd for their helpful feedback on the paper. We thank Maren Smith from the BYU Law School for assistance with the Westlaw queries and court case analysis. We also thank Joshua Reynolds and Scott Ruoti for their feedback on an earlier draft of the paper. This material is based upon work supported by the National Science Foundation under Grant No. CNS-1816929, by the KU Leuven Research Council under grant C24/18/049, by CyberSecurity Research Flanders with reference number VR20192203, and by DARPA FA8750-19-C-0502.

References

[1] Vincent Bindschaedler, Reza Shokri, and Carl A Gunter. Plausible deniability for privacy-preserving data synthesis. *arXiv preprint arXiv:1708.07975*, 2017.

[2] Nikita Borisov, Ian Goldberg, and Eric Brewer. Off-the-record communication, or, why not to use PGP. In *Proceedings of the Third ACM Workshop on Privacy in the Electronic Society (WPES)*, 2004.

[3] Ekaterina Damer and Phelim Bradley. Prolific: A platform for large-scale data collection and processing. https://www.prolific.co/. Accessed: 2023-01-21.

[4] Mario Di Raimondo, Rosario Gennaro, and Hugo Krawczyk. Deniable authentication and key exchange. In *Proceedings of the 13th ACM Conference on Computer and Communications Security (CCS)*, 2006.

[5] S. Dixon. Most popular global mobile messenger apps as of January 2023. https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/.

[6] Satu Elo and Helvi Kyngäs. The qualitative content analysis process. *Journal of advanced nursing*, 62(1):107–115, 2008.

[7] Asghar Ghasemi and Saleh Zahediasl. Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, 10(2):486, 2012.

[8] Matthew Green. Ok Google: please publish your DKIM secret keys. https://blog.cryptographyengineering.com/2020/11/16/ok-google-please-publish-your-dkim-secret-keys/. Accessed: 2023-01-21.

[9] Mike Hearn. On the value of deniability in OTR like protocols. https://moderncrypto.org/mail-archive/messaging/2014/001173.html.

[10] Ole R Holsti. Content analysis for the social sciences and humanities. *Reading. MA: Addison-Wesley (content analysis)*, 1969.

[11] Michal Kedziora, Yang-Wai Chow, and Willy Susilo. Threat models for analyzing plausible deniability of deniable file systems. 2017.

[12] Christiane Kuhn, Maximilian Noppel, Christian Wressnegger, and Thorsten Strufe. Plausible deniability for anonymous communication. In *Proceedings of the 20th ACM Workshop on Privacy in the Electronic Society (WPES)*, 2021.

[13] Sam Liang and Yun Fu. Otter: A company that develops speech-to-text transcription applications. https://www.otter.ai. Accessed: 2023-01-21.

[14] Loveday Morris. Andrew Tate rape allegations, text exchanges with women detailed in court document. https://www.washingtonpost.com/world/2023/02/02/andrew-tate-rape-investigation-romania.

[15] NIST. Non-repudiation. https://csrc.nist.gov/glossary/term/non_repudiation.

[16] Quirkos. A qualitative analysis software. https://www.quirkos.com/. Accessed: 2023-01-21.

[17] Blake Ramsdell and Sean Turner. Secure/Multipurpose Internet Mail Extensions (S/MIME) version 3.2 message specification. Technical report, 2010.

[18] Nathan Reitinger, Nathan Malkin, Omer Akgul, Michelle L Mazurek, and Ian Miers. Is Cryptographic Deniability Sufficient? Non-Expert Perceptions of Deniability in Secure Messaging. In *IEEE Symposium on Security and Privacy (SP)*, 2023.

[19] Thomson Reuters. Westlaw: A legal research platform. https://legal.thomsonreuters.com/en/westlaw.

[20] Michael Schliep and Nicholas Hopper. End-to-end secure mobile group messaging with conversation integrity and deniability. In *Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society (WPES)*, 2019.

[21] Michael A Specter, Sunoo Park, and Matthew Green. KeyForge: Non-Attributable Email from Forward-Forgeable Signatures. In *USENIX Security Symposium*, 2021.

[22] Tor. Tor project. https://www.torproject.org/.

[23] Nik Unger and Ian Goldberg. Improved strongly deniable authenticated key exchanges for secure messaging. *Proceedings on Privacy Enhancing Technologies (PoPETS)*, 2018(1):21–66, 2018.

[24] Nihal Vatandas, Rosario Gennaro, Bertrand Ithurburn, and Hugo Krawczyk. On the cryptographic deniability of the signal protocol. In *International Conference on Applied Cryptography and Network Security (ACNS)*. Springer, 2020.

[25] Juan Vera-del Campo, Josep Pegueroles, Juan Hernández-Serrano, and Miguel Soriano. Doccloud: A document recommender system on cloud computing with plausible deniability. *Information Sciences*, 258:387–402, 2014.

[26] Philip R Zimmermann. *The official PGP user's guide*. MIT press, 1995.

A Appendix

A.1 Westlaw case queries

In Section 6, we mentioned that we use Westlaw [19], an online legal research service and proprietary database for lawyers and legal professionals, to retrieve court cases (judicial opinions). We now provide details on how we created

Westlaw search queries that return the cases where WhatsApp chat was potentially considered admissible evidence and where it was potentially rejected as evidence.

Heuristics to find court cases where WhatsApp is potentially accepted as evidence:

- The word "WhatsApp" is in the same paragraph as one of the following words: evidence, admissib!, foundation, authenticat!, relevan!, accept! or support!. Here ! searches for words with multiple endings.
- There exists a paragraph where the word "hearsay" is less than five words away from either of the following words "no," "not," and "exception." Also, this paragraph has the word "WhatsApp" somewhere in it.
- Exclude court cases where "WhatsApp Inc." is in the title.

WhatsApp potentially accepted as evidence: *WhatsApp /p evidence admissib! (no not exception /5 hearsay) foundation authenticat! relevan! accept! support! % TI(WhatsApp Inc.)*

Heuristic to find court cases where WhatsApp is potentially rejected as evidence:

- The word "WhatsApp" is in the same paragraph as at least one of the following words: forge, deny, spoliation, fail, neglect, invalid, deniab!, inadmissib!, unaccept!. Here ! searches for words with multiple endings.
- There exists a paragraph where one of these words (no, not, lack!, fail!, neglect!) is less than five words away from either of the following words (admissib!, foundation, authenticat!, relevan!, evidence) Also, this paragraph has the word "WhatsApp" somewhere in it.
- There exists a paragraph where one of these words (limine, object!, exclud!) is less than five words away from either of the following words (grant!, evidence). Also, this paragraph has the word "WhatsApp" somewhere in it.
- Exclude court cases where "WhatsApp Inc." is in the title.

WhatsApp potentially rejected as evidence: *WhatsApp /p forge deny spoliation deniab! fail neglect inadmissib! unaccept! invalid (no not lack! fail! neglect! /5 admissib! foundation authenticat! relevan! evidence) (limine object! exclud! /5 grant! evidence) % TI(WhatsApp Inc.)*

A.2 Semi-Structured Interview Script

Thank you for participating in our interview today. Before we get started we will need you to read through a consent form and confirm that you are willing to participate in our study. I am sending you a link to the form in the chat. Let me know when you finish reading and are ready to proceed.

This interview aims to know your views, opinions, and understanding regarding Internet communication (e.g., email, instant messaging apps). While answering our questions, please keep in mind that there is no single correct answer to these questions. Please answer the questions based on your knowledge and experiences.

Understanding of digital communication

- What type of Internet applications do you use to communicate with other people over the Internet?
- I'm going to ask you to explain your perceptions and ideas about communication between two people over the Internet. This is a drawing exercise. Assume you receive a message over Email. Draw a diagram showing the entities involved and the message's path from the sender to you. Please talk aloud and explain your thought processes while you are drawing. *Ask about who can read or modify the messages during transit.*
- *Repeat the process for messaging applications such as Facebook Messenger, Signal app, and WhatsApp.*
- *Repeat the process for text messaging.*

Verification of the sender's identity Assume you receive a message over email, Facebook, or another messaging application. How do you confirm the sender's identity?

Proving the sender's identity to others

- Assume you receive a message over email, Facebook, or another messaging application. How can you use this message to prove to others that the sender sent you this message?
- Will your answer vary if you have to prove the sender's identity to (1) your family, (2) friends, (3) social media connections, or (4) in a court?
- Will your answer vary if the message you need to prove is from (1) your family, (2) friends, (3) social media connections, or (4) in a court?

Verifying the claim from others

- Imagine you have two friends who are also friends with each other. If friend A tells you that friend B was talking bad about you behind your back, would you believe them? Why or why not? What would you do if you wanted to verify it? How would it affect your relationship with friends A or B?
- Assume the same situation, but now friend A also shows you messages from friend B that say bad things about you. How would your reaction be different from the previous scenario?
- How would your answer vary based on different scenarios such as the content of the message, your relationship with person A who is claiming, and your relationship with person B? (1) Person A/B connection: friend, family, social media connection, acquaintance, stranger.(2) Content of the message: bad things about you, bad things about other friends, bad things about himself (person B), etc.

Understanding of OTR's deniability definition

- Now I will present a definition of a feature that any online communication application such as email or Facebook messenger could provide. Please explain your understanding of this definition and what you get from this

feature as a user. *The messages you send do not have digital signatures that are checkable by a third party. Anyone can forge messages after a conversation to make them look like they came from you. However, during a conversation, your correspondent is assured the messages he sees are authentic and unmodified.*

- To your knowledge, are there any apps or websites that already provide this feature?
- *If yes*, do you think it's beneficial, or do you find some pitfalls in this?
  If not, do you think it would be beneficial if some apps incorporate this?

Preference for authentication properties

- Now I am going to send you definition of three different features that can be implemented by any application/website that allows you to communicate with people, such as email or messaging apps. Choose among these three features that you would like to use for your internet apps or websites. Also, provide reasoning for your choice?
  - Indisputable: Every message sent (or received) over the Internet would have the sender's (or receiver's) identity bound to it. Thus, the recipient of your messages can prove to anyone else that you sent the message.
  - Disputable: Any message sent or received over the Internet would not have the sender's identity bound to the message. However, the recipient of your messages can still verify that the message came from you but cannot prove to anyone else that you sent the message.
  - Anonymity: Any message sent or received over the Internet would not have the sender's identity bound to the message. But in this case, even the recipient of a message does not know the sender's identity.
- Assume you can have a combination of the above features based on various scenarios. Would you choose only one feature for all scenarios and applications or a combination of these properties? What combination you would like to have for your Internet communication?

A.3 Survey Questions

The exact representation of questions is available at https://bitbucket.org/isrlauth/deniability/src/master/survey.pdf.

Please enter the first name or nicknames of people you know for each category below. We will use these names to personalize the survey questions for you.

Please ensure you remember the category for each person you enter. We will be using the names you enter to ask further questions.

The responses to this question are only used to personalize your survey. We ensure that only researchers have access to the dataset. Your responses to this question will be deleted within three weeks after the survey. *Participants see a 3 rows x 2 cols table, with the following rows: Family/ Close friends (most trusted), Acquaintances (somewhat trusted), Less trustworthy persons, and with following columns Name 1, Name 2.*

Question 1 Assume three scenarios where person A tells you that person B was talking bad about you behind your back.

Scenario 1: Person A tells you verbally.
Scenario 2: Person A shows you a screenshot of the conversation with person B that happened in their messaging app (e.g., WhatsApp). In the screenshot, person B was talking bad about you.
Scenario 3: Person A shows you messages from person B in their messaging app (e.g., WhatsApp) in person. In the chat, person B was talking bad about you.
How much do you trust person A's claim? Fill out your trust level on a scale of 1 to 10, where 1 is no trust and 10 is complete trust.
participants see a 9 rows x 3 cols table, where each row represents a combination of Name 1 and Name 2, such as "Name1 tells you that Name2 was talking bad about you behind your back.". The columns were Verbal, Shows you Screenshot, Shows Chat in messaging app. Participants need to fill in trust for all 9x3 boxes.

Question 2: Assume you went for dinner with a group last month. From that group, Person A tells you that you owe them $50 for dinner. However, you don't remember whether they paid for your dinner. Assuming three possible scenarios
(1): Person A tells you verbally that you owe them $50.
(2): Person A shows you a screenshot of the conversation in their messaging app. In the chat, you acknowledged that you owed them $50.
(3): Person A shows you the messaging app's chat in person. In the chat, you acknowledged that you owed them $50.

How likely will you trust their claim in each scenario and give them the $50? Fill out your trust level on a scale of 1 to 10, where 1 is no trust and 10 is complete trust. *participants see a 3 rows x 3 cols table, where each row represents each of Name 1. The columns were Verbal, Shows you Screenshot, Shows Chat in messaging app. Participants need to fill in trust for all 3x3 boxes.*

Question 2b: Please select in which category each person exists. *Participants see all 6 names from name 1 and Name 2 categories and they need to identify the correct category for each of the names from (1) Family/ close friend, (2)Acquaintance, and (3) Less trustworthy person.*

Question 3: You recently contacted a landlord about renting an apartment. They offer you a $200 Amazon coupon if you sign the contract within two days. If you want this coupon, which of the following option(s) would you agree to:

- You will sign the contract the same day and trust the landlord to deliver the coupon as promised.

- You will sign the contract the same day after the landlord sends you the offer details on the messaging application, assuming you can later use the chat as evidence (in case of any disputes).
- You will first ask the landlord to add the offer to the contract. Then, you will wait to sign until they add it to the agreement. This could delay the contract signing process by at least a day.
- Others:

Question 4: You moved to a new apartment and found that the air conditioner is damaged. Although you do not need it, you want to ensure that the landlord does not blame you later for it. Which of the following options would you be willing to use?

- You will inform your landlord in person.
- You will inform your landlord on a phone call.
- You will send them a message on a messaging application about the damage, assuming you can later use the chat as evidence (in case of any disputes).
- You will ask the landlord to edit the contract to include damage details.
- Others:

Question 5a: Give an example where you needed to prove to others that person X sent you a message using online communication like messaging apps, email etc.

Question 5b: In your previous response, how important was it for you to prove to others that Person X indeed sent the message?

Question 6a: Give an example where you used online communication (like messaging apps or email) and required that the recipient is sure that you sent the message, but they cannot prove to others that you are the original sender.

Question 6b: In your previous response, how important it was for you that the recipient cannot prove to others that you were the original sender?

Question 7a: Assume you send a message to person X. Suppose they need to prove to a third party that you sent them that message. Which of the following option(s) are enough to convince a third party?

- Person X can show screenshots to the third party.
- Person X can forward the message to the third party.
- Person X can show the chat in the messaging app (in person) to the third party.
- Person X cannot prove to a third party that you sent the message.

Question 8a: Assume there is a messaging app that provides the following "Deniability" property:
Definition: The messages you send do not have digital signatures that are checkable by a third party. Anyone can forge messages after a conversation to make them look like they came from you. However, during a conversation, your correspondent is assured the messages he sees are authentic and unmodified.
Please select a choice for each of the following statements.

- You completely understand the definition
- Definition is self-contradictory
- An attacker can send you a message, impersonating your friend
- The messaging server or a third party cannot determine the sender's identity
- A third party can modify the content of messages sent by your friend
- The recipient cannot prove to a third party that they received a message from the sender
- The recipient does not know the sender's identity

Question 8b: According to your understanding, which of the following apps provide Deniability?
Deniability: The messages you send do not have digital signatures that are checkable by a third party. Anyone can forge messages after a conversation to make them look like they came from you. However, during a conversation, your correspondent is assured the messages he sees are authentic and unmodified. *Display the following rows WhatsApp, Signal app, Facebook messenger, Texting (SMS), Gmail, Yahoo Mail and following columns yes, no, not sure.*

Question 9a: Select all the properties from the list that you want for your Internet communications: *displays the images shown in Figure 4* Indisputable: Every message sent (or received) over the Internet would have the sender's (or receiver's) identity bound to it. Thus, the recipient of your messages can prove to anyone else that you sent the message.

Disputable: Any message sent or received over the Internet would not have the sender's identity bound with the message. However, the recipient of your messages can still verify that the message came from you but cannot prove to anyone else that you sent the message.

Anonymous: Any message sent or received over the Internet would not have the sender's identity bound to the message. But in this case, even the recipient of a message does not know the sender's identity.

Question 9b: Since you selected multiple properties in the previous question, how do you expect to get them over the Internet? Below are the same definitions for your reference: *Description of Message-based, Application-based, time-based, Duration-based, Person-based, and an option for Others.*

Question 9c: Assuming you can get only one property per application, which property would you choose for the following applications? Below are the same definitions for your reference: *display following rows - Messenger apps (e.g., Facebook messenger), SMS, Email*

Question 10a: How hard is it to forge a screenshot of a chat?

Question 10b: How hard is it to forge a message in a messaging application?