Is Cryptographic Deniability Sufficient?
Non-Expert Perceptions of Deniability in Secure Messaging

Nathan Reitinger, Nathan Malkin, Omer Akgul, Michelle L. Mazurek, and Ian Miers
University of Maryland

Abstract—**Cryptographers have long been concerned with secure messaging protocols threatening deniability. Many messaging protocols—including, surprisingly, modern email—contain digital signatures which definitively tie the author to their message. If stolen or leaked, these signatures make it impossible to deny authorship. As illustrated by events surrounding leaks from Hilary Clinton's 2016 U.S. presidential campaign, this concern has proven well founded. Deniable protocols are meant to avoid this very outcome, letting politicians and dissidents alike safely disavow authorship. Despite being deployed on billions of devices in Signal and WhatsApp, the effectiveness of such protocols in convincing people remains unstudied. While the absence of cryptographic evidence is clearly necessary for an effective denial, is it sufficient?**

We conduct a survey study ($n = 1,200$) to understand how people perceive evidence of deniability related to encrypted messaging protocols. Surprisingly, in a world of "fake news" and Photoshop, we find that simple denials of message authorship, when presented in a courtroom setting without supporting evidence, are not effective. In contrast, participants who were given access to a screenshot forgery tool or even told one exists were much more likely to believe a denial. Similarly, but to a lesser degree, we find an expert cryptographer's assertion that there is no evidence is also effective.

1. Introduction

Cryptographic deniability—a problem long considered largely theoretical—has recently come to play a major role in world events. In the closing weeks of the 2016 United States presidential election, approximately 58,000 emails from Hillary Clinton's campaign were publicly leaked [1], [2]. The Clinton campaign broadly denied the authenticity of the emails, claiming that they were doctored as part of a smear campaign [3]–[5]. Since emails are generally unauthenticated, they give plausible *deniability*: conversation participants can claim they did not author a message. Unfortunately for the Clinton campaign, security researchers soon pointed out a problem with the campaign's denials: the emails were cryptographically signed—not by the authors—but by the Mail Transfer Agents' (e.g., Google's servers) use of Domain Keys Identified Mail (DKIM), a security feature implemented by many email providers to combat spam [6]. The messages were therefore verifiably unaltered [7]–[9]. Cryptographic signatures on the emails rendered the denials ineffective, a pattern that has repeated in subsequent incidents as recently as March 2022 [10].

Political emails are one highly salient situation where the ability to deny message authorship can be valuable, but not of course the only one; deniability has also been proposed, for example, as a tool for dissidents to avoid persecution by denying authorship of heretical messages.

Cryptographic deniability In 2004, Borisov, Goldberg, and Brewer proposed Off-the-Record (OTR) [11], a protocol for encrypting and authenticating text messages while removing the non-repudiation property provided by then-standard signature-based approaches for authenticated encrypted communication such as GPG and S/MIME [12]–[14], and consequently enabling deniability. OTR and similar protocols allow participants in an encrypted chat to know messages are authentic during the chat, but provide no way to transfer this knowledge to anyone else: there are no signatures that a third party can check.[1] Since then, deniability in encryption protocols has seen considerable activity in both academia and industry [15]–[17]. Successors to OTR are now deployed in over two billion devices worldwide through services like WhatsApp and Signal [18], [19].

Cryptographic deniability protocols like OTR implicitly make a strong assumption: the absence of signatures on messages is both necessary and *sufficient* for deniability, i.e., unsigned messages, when revealed to third parties or the public, will not be trusted by anyone. While this is the appropriate scope for the technical challenge of developing deniable cryptographic protocols, it leaves open an important question: given a cryptographically deniable messaging protocol, what else, if anything, does it take to achieve deniability in practice?

To our knowledge, however, the question of how cryptographic deniability interacts with human perceptions—and consequently, how to improve real-world deniability—has not previously been studied.

Research questions Are human decisions governed by the same logic as cryptographic deniability? When faced

[1] In OTR, Alice and Bob share a symmetric key used to generate authentication tags for messages. Alice can regenerate the tag on received messages and, if it matches, conclude that Bob authored the message (since only she and Bob know the key and she didn't write the message). A third party, however, cannot verify the tag without the key. For added deniability, OTR periodically rekeys and publishes the old key, allowing anyone to make a forgery. Subsequent work uses stronger methods [15].

with a text message, an alleged author, and a denial, how do people decide who to believe and how much are they willing to act on those beliefs? As we saw in the 2016 U.S. presidential election, this is no longer a hypothetical.

If the absence of cryptographic evidence is necessary for deniability, but *insufficient*, what kind of evidence will people look for? What will be their default assumptions? And what does it take for a protocol to achieve not just cryptographic deniability, but *human* deniability that convinces actual people a message may have been forged? These are the questions motivating our work.

More specifically, as a first step, our study sought to address the following research questions:

RQ1: Are people more likely to disbelieve alleged messages when the denial is accompanied by supporting evidence?

RQ2: Are some types of supporting evidence for denials more effective at changing beliefs than others?

RQ3: Does supporting evidence affect not just belief in message denial, but also (intended) decisions related to that belief?

RQ4: How do individual attributes like gullibility or political leanings affect beliefs about message denial?

Approach To address our research questions, we conducted a survey study with 1,200 people that presented various arguments for deniability to different participants and gauged their reactions. To remove as many potential confounds as possible, but retain a context that would feel realistic for our participants, the scenario for our survey was a criminal trial. Participants played the role of jury members and were asked to judge the guilt of a hypothetical politician, who is accused of accepting a bribe. The evidence for the bribery charge is a screenshot from the politician's messaging history. We will discuss the setup in greater detail in Section 3.

As potential defenses, participants evaluated one of six different forms of deniability evidence. This deniability evidence fell into three general categories: no supporting evidence (baseline), experts who testify about the properties of the protocol, and the demonstration of tools that make transcript forgery trivial. After participants reviewed this evidence, we asked them about their beliefs, and then to make a decision based on the case presented.

Key results We offer two key findings. First, one might assume that familiarity with "photoshopped" images, so-called "fake news," and even "deepfake" videos[2] would all support the cryptographic perspective on deniability: anything that cannot be affirmatively authenticated should be assumed to be fake [20]–[23]. To our surprise, however, even though 70% of participants either had or believed they could fake a screenshot, two-thirds of participants who saw only an assertion (without further evidence, see Section 3.1) that the message was fake believed that a simple screenshot convincingly demonstrated wrongdoing. In contrast, no more than 26% of participants who saw any type of deniability evidence were convinced of wrongdoing.

Second, we find that some types of evidence are more convincing than others: participants who tried out an interactive forgery tool were significantly more likely to believe a denial than those who made their decisions on the basis of testimony from expert cryptographers. Overall, we find that deniability is obtainable, but messaging applications can do more to help make deniability effective in practice.

2. Background and Related Work

Deniability is far from a new concept. Although the inclusion of deniability as a feature of secure messaging applications is relatively recent, deniability has a rich, non-technical history, giving us insight into what deniability might mean to users—i.e., how deniability evidence should be designed. The following section provides a brief overview of deniability from a non-technical perspective, followed by a look at how deniability has been developed as a feature in secure messaging protocols, and then how usable or unusable those protocols are.

2.1. Deniability in philosophy and politics

Philosophical concepts of deniability focus on *ambiguity*, using a statement-by-statement setting to analyze the concept. In philosophy, researchers define deniability as something to be attained through plausible explanations: deniability is gained when an utterance has multiple meanings, meaning that the speaker may deny having meant *one* of the particular meanings over another [24]–[27].[3]

From a U.S. political perspective, deniability matured during the Cold War [29]. Again, *ambiguity* is a key animating concept, as described in the National Security Council's 1948 definition of covert actions: "[Operations] so planned and executed that any US Government responsibility for them is not evident to unauthorized persons and that if uncovered the US Government can plausibly disclaim any responsibility for them [30]." This approach is perhaps best exemplified by the Iran-Contra affair of the 1980s. During this time period, the United States was rumored to be involved in, and later found responsible for, the sale of arms to Iran, during an arms embargo, prospectively to fund the Contras [31]. Although President Reagan was implicated in these plans, he was able to believably deny any knowledge of the scheme [32]. Skepticism regarding the president's involvement was validated by a congressional investigation following the scandal, but the investigation found the President responsible only in a should-have-known sense [33].

[2] For example, in a 2016 Pew study, 88% of American adults believed fake news was causing some or a great deal of confusion [20], and about half of college students in a 2021 survey were aware of deepfakes [21].

[3] "Matt is running out of fuel and needs some fast. He stops and asks a stranger where he can get some fuel. The stranger says 'there is a gas station around the corner.' The stranger thereby implies (implicitly communicates) that the gas station is open and has fuel" [27]. If, however, the gas station were not open, the stranger could state he merely provided an option, and did not say that the gas station were open or had gas; the stranger's statement has multiple interpretations due to ambiguity [28].

2.2. Deniability in cryptographic protocols

Since deniability first appeared in computer security, there has been no universal agreement on its formal definition [34]. That said, one generally accepted cryptographic definition is stated in Unger et al.'s 2015 systematization of knowledge paper: "Given a conversation transcript and all cryptographic keys, there is no evidence that a given message was authored by any particular user" [16], [35]. In simpler terms, lack of cryptographic proof of authorship—and the resulting ambiguity—provides deniability.

Deniability in secure communications protocols can be divided into two types: participant deniability, or the denial of participation in an entire conversation, and message deniability, or the denial of one or more individual messages [16]. Here, we assume deniable schemes provide both types. This both simplifies the denial story for study participants and is in line with the deniability features of existing encrypted messaging apps like Signal and WhatsApp.

The first practical, deniable secure messaging scheme, OTR, achieves deniability with a Deniable Authenticated Key Exchange (DAKE) plus malleable encryption [11], [36]. A shared ephemeral key created with a long-term, private key allows sending and receiving parties to authenticate communications with each other. Shared ephemeral keys and malleable encryption allow recipients to manipulate incoming messages, making them indistinguishable from unaltered messages. If a ciphertext can be meaningfully altered by a recipient, then any messages purported to be authored by the sender could have theoretically been created instead by the recipient, providing deniability.

Signal's encrypted messaging protocol—also deployed in other messaging applications like WhatsApp—has become the de facto standard for secure messaging today. The Signal protocol improves on OTR's deniability by modifying its DAKE [37]. In OTR, forging a transcript between Alice and Bob requires either Alice or Bob's private key or a transcript of a legitimate conversation between Alice and Bob to edit. As a result, forgeries can only practically be fabricated by conversation participants themselves, not any third party.

To achieve broader deniability, Signal uses an Extended Triple Diffie-Hellman (X3DH [38]) key agreement that computes a shared secret derived from multiple Diffie-Hellman key exchanges: (1) between the sender's and recipient's short-term keys; (2) between the sender's long-term key and the recipient's short-term key; and (3) between the sender's short-term key and the recipient's long-term key. More details, which may subtly impact deniability, may be found in the protocol specifications [38]. Because the shared secret can be computed with knowledge only of both ephemeral secrets, any party can forge a transcript between two long-lived public keys [15], [18].[4]

Cryptographic deniability has also been studied under specific constraints like email. Specter, Park, and Green devised a scheme which provides non-attributable DKIM signatures for email [39] by releasing signing keys and allowing retroactive undetectable forgeries. As a result, signed emails leaked at a later date may be genuine or they may be forgeries, and there is no way to to tell. Beck et al. use a similar concept of time-deniable signatures, but provide time-based computation limitations to enforce timing restraints [40].

2.3. Secure messaging usability

Although deniability, specifically, has not been widely studied in terms of usability or human perceptions, secure messaging in general has. From the first "Why Johnny Can't Encrypt" study to the more recent "When SIGNAL hits the Fan," researchers have identified gaps in understanding that can (and often do) undermine effective security [41], [42]. For example, Tan et al. found that key verification used by applications like OTR and WhatsApp was insufficient to protect users from a man-in-the-middle attack [43]. Similar findings have applied to applications like Signal, Facebook Messenger, Telegram, and many other secure messaging tools [42], [44]–[48].

Additionally, researchers have identified gaps between developers' and end users' beliefs about security [47], [49]–[51]. Ermoshina et al. conducted international interviews with high- and low-risk users, and with developers, focusing on security concepts found in secure messaging tools [51]. Although the researchers did not focus heavily on deniability, they did find that for some high-risk users, ephemeral messages were of higher importance than cryptographic deniability, and that some users had developed ad-hoc practices to try to achieve ephemerality and related goals. This motivates our examination of how well cryptographic deniability works in practice and how it might be improved.

3. Methods

To assess people's perspectives on deniability, we designed a between-subjects survey study which presented different kinds of deniability evidence and then asked questions about participants' beliefs. We initially ran the study with 600 participants (**Survey 1**, December 2021). After analyzing results, we conducted a second study **Survey 2** ($n = 600$, February 2022), with greater statistical power, to examine a subset of the effects in detail.

3.1. Study design

One of the major challenges of this research was developing a study design that would isolate perceptions of deniability from any potential confounds while retaining as much ecological validity as possible.

Pilot reveals design challenges Our initial survey design used a scenario involving the press leak of a politician's private messages, mirroring high-profile leaks in the real

[4] In contrast, in a real conversation, Alice knows she generated one of the ephemeral keys and her long-term key and knows she kept the private keys secret, so she knows the conversation is authentic.

world [52], [53]. We piloted this version with four participants, asking them to think aloud. The pilot participants struggled to decide whether or not they believed the denials, in large part because they wanted more context: without more evidence about the politician in question, their history, and the overall political environment, the pilot participants were unable or unwilling to evaluate specific deniability arguments.

After the pilots, we revisited the study scenario. We considered adding the context the pilot participants requested, which would add more ecological validity, but would also create several critical problems. First, the space of contextual factors is unmanageably large, and they are hard to disentangle: everything—a subject's track record, their politics, their status, their appearance—could contribute to the persuasiveness of an argument. By adding context, we may also motivate participants to (dis)believe the denial on the basis of the participant's political affiliation or background. Further, these contextual factors are external to messaging apps themselves; learning how context affects deniability will not necessarily lead to concrete recommendations for protocol designers.

As a starting point for understanding the impact of cryptographic deniability, we wanted to isolate only the messaging properties in order to understand their direct effects; we expect that follow-up work will build on our findings to place these results in more context.

Courtroom setting To resolve this dilemma, we searched for an approach that would let us isolate concrete factors that would be actionable for protocol designers, yet would not distract participants due to a lack of context. We identified a solution that we believe offers a practical compromise: setting our survey in a courtroom. Trials are one of the few places where people are used to being told to disregard bias in favor of specific pieces of evidence [54]. We therefore felt that asking participants to play the role of a jury member would provide a convincing explanation for why they were being asked to make decisions on the basis of very limited evidence, rather than a broader set of facts.

Concretely, we asked participants to role-play that they were jurors in a bribery trial of a governor, where the key piece of accusatory evidence was a screenshot of a messaging app; the defense presented different forms of deniability evidence in order to argue for the governor's innocence. We tested this jury-trial framing with a small number of participants from the same recruitment pool as the main study (Section 3.4); finding it successful, we included those initial participants in our sample and proceeded to full recruitment.

Dependent variable Moving to the courtroom framing opened up two possibilities for measuring the primary dependent variable of our study: whether a respondent finds a particular piece of deniability evidence convincing. One is to ask about a participant's belief and the strength of that belief; the other, more trial-specific approach is to ask about which verdict they would render ("guilty" or "not guilty").

We decided to ask both forms of this question, because there were no clear grounds for preferring one over the other, and we had a basis to believe that people may approach them in different ways. Namely, a verdict represents an action (voting), and a literature review grounded in opinion dynamics and attitude-behavior consistency/inconsistency (social psychology) found that there may be more confidence required to act versus hold a belief [55]–[57].

Discrepancies between beliefs and verdicts could also occur if people have a higher bar for voting "guilty" compared with holding a belief in someone's guilt. This represents a potential limitation of the courtroom setting, which we discuss next.

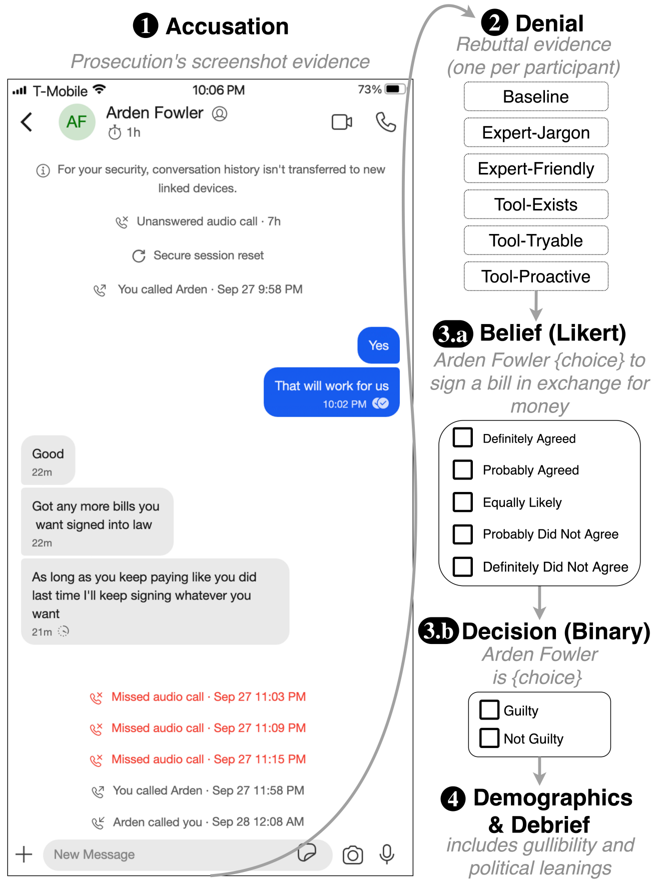
Courtroom biases While framing our study around a trial allows us to sidestep many potential confounds due to contextual factors, it potentially introduces its own set of unique biases given that courtrooms are highly evidence oriented. While we purposefully did not instruct participants about evidentiary standards, media coverage or personal trial experiences may expose people to various legal ideas, which, in turn, can influence their decisions. For example, someone who internalized the notion of "proof beyond a reasonable doubt" may require a higher bar of proof in a courtroom scenario than in other scenarios. Similarly, those familiar with the legal principle "innocent until proven guilty" may lean towards acquittal if they are uncertain. On the other hand, some may hold the opposite position and view the mere fact of a trial as evidence of culpability, believing that the prosecution would not bring a case without substantive reasons.

As the goal of our study was to compare different types of deniability evidence, we presented each in their own between-subjects conditions (Section 3.3). The biases discussed so far affect every type of evidence equally, enabling comparisons.

Baseline condition As in most between-subjects experiments, our study includes a baseline condition that we use as a point of comparison, but our courtroom scenario presents some challenges in designing it. We chose to have the defendant simply state that the message is fake, with no evidence (see Section 3.3). In a real courtroom, a defense with no evidence could be inherently suspicious, and therefore not as neutral as we might prefer in a baseline. Other baseline options may also have been valid; however, we were unable to identify an alternative that would be truly neutral. Though a simple denial may be inherently suspicious, complex denials rely on context or add other kinds of evidence, which create challenges for comparison.

Because we focus primarily on direction of effect ('*is* this more convincing, compared with no evidence?') rather than on magnitude ('how much more convincing is this, compared with no evidence?'), we believe our choice of baseline is reasonable, despite this neutrality challenge.

Generalizability This section has discussed the various biases of the courtroom setting, which raises the question of ecological validity: will people in the real world make



T-Mobile 10:06 PM 73%

AF Arden Fowler
1h

For your security, conversation history isn't transferred to new linked devices.

Unanswered audio call · 7h

Secure session reset

You called Arden · Sep 27 9:58 PM

Yes

That will work for us
10:02 PM

Good
22m

Got any more bills you want signed into law
22m

As long as you keep paying like you did last time I'll keep signing whatever you want
21m

Missed audio call · Sep 27 11:03 PM

Missed audio call · Sep 27 11:09 PM

Missed audio call · Sep 27 11:15 PM

You called Arden · Sep 27 11:58 PM

Arden called you · Sep 28 12:08 AM

New Message

Figure 1: Participants were shown the prosecution's screenshot evidence ❶ followed by one version of the defense's rebuttal ❷. Afterwards, participants were asked a five-point question about belief 3.a and to make a binary decision about guilt 3.b. Lastly, we asked demographic questions ❹, including gullibility and political leanings, and closed with a debriefing.

decisions about deniability in the same way as the "jurors" in our study? Likely, no; in real life, no evidence will ever exist in isolation. However, the factors we examine here will play some part in real-life decisions, and by isolating them we are able to understand them more precisely.

Like any study, ours involves trade-offs. By zooming in on people's decisions in a way that sacrifices some realism, we are able to obtain a more detailed idea of what to look at in follow-ups that will have greater ecological validity. This is necessary because our study is, to our knowledge, the first work on this topic; by constructing it in this way, we provide a foundation for future work to build on.

3.2. Survey protocol

Overview The survey protocol consisted of four main sections: (1) a description of the framing scenario; (2) the deniability evidence; (3) questions about participants' interpretation of the evidence; and (4) demographic questions and a debriefing. An overview of the study flow may be seen in Figure 1. The full survey is available in Appendix A.

In the first section of the survey, following informed consent, we asked participants to imagine a scenario where they were selected to serve on a jury during a trial (see Section 3.1 for further explanation regarding the choice of a courtroom). Participants first saw a piece of accusatory evidence referred to by the prosecution as "The Leak"—a text message screenshot supposedly authored by a local politician, Governor Arden Fowler. Based on statements found in the leaked screenshot, the prosecution argues that Governor Fowler is guilty of bribery, a federal crime [58]. Following the accusatory screenshot, participants answered comprehension questions to ensure an understanding of the screenshot's implication regarding bribery.

Next, in the second section, participants saw one of six types of evidence potentially providing deniability (detailed in Section 3.3), presented as the defense's rebuttal to the accusation. To allow for between-subjects comparisons, we kept all other parts of the survey consistent, but varied the type of deniability evidence individual participants saw, allowing us to measure the effectiveness of the evidence by comparing responses in different experimental conditions. Participants were again asked a comprehension question about the evidence shown.

In the third section, we asked about participants' belief that Governor Fowler accepted a bribe (five-point Likert-type scale) and for their verdict choice (binary). We also asked some free-response questions about the evidence they saw, as well as what other evidence might change their mind.

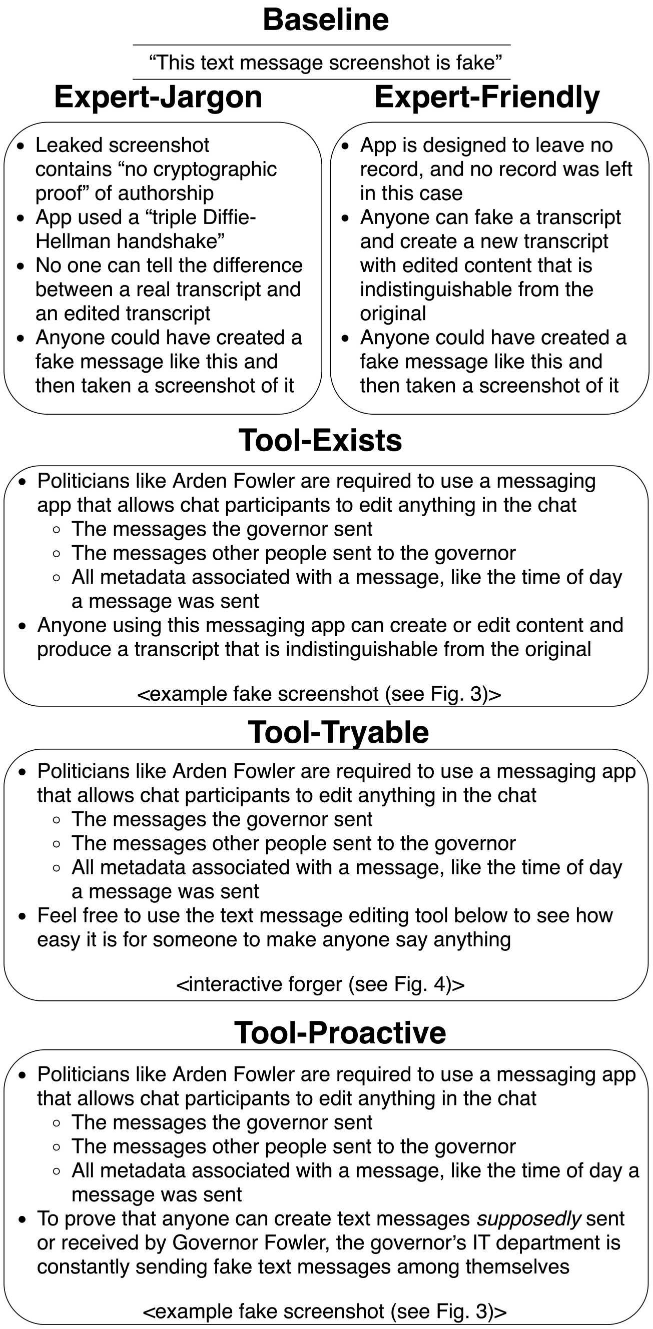
The final section of the survey consisted of a 12-item gullibility scale [59], a political leanings question (five points from very conservative to very liberal), a question about experience with fake screenshots, and several demographic questions (age, gender, education, and technology literacy). Participants were then debriefed and informed that the study was designed to determine how users evaluate the authenticity of text messages in the context of "leaked" screenshots.

3.3. Conditions

To measure the effectiveness of various deniability arguments, our survey assigned each participant to one of six experimental conditions, which determined the deniability evidence shown to the participant: BASELINE, EXPERT-JARGON, EXPERT-FRIENDLY, TOOL-EXISTS, TOOL-TRYABLE, and TOOL-PROACTIVE. The first condition acts as a baseline, the next two conditions rely on experts, and the final three conditions involve the existence of tools that make screenshot forgery trivial. A summary of each condition is given in Figure 2; the full text can be found in Appendix A. The following subsections describe each condition in turn.

3.3.1. Baseline.

The BASELINE condition provides the weakest possible rebuttal evidence. The defense simply asserts: "This text message screenshot is fake."



Baseline

"This text message screenshot is fake"

Expert-Jargon

- Leaked screenshot contains "no cryptographic proof" of authorship
- App used a "triple Diffie-Hellman handshake"
- No one can tell the difference between a real transcript and an edited transcript
- Anyone could have created a fake message like this and then taken a screenshot of it

Expert-Friendly

- App is designed to leave no record, and no record was left in this case
- Anyone can fake a transcript and create a new transcript with edited content that is indistinguishable from the original
- Anyone could have created a fake message like this and then taken a screenshot of it

Tool-Exists

- Politicians like Arden Fowler are required to use a messaging app that allows chat participants to edit anything in the chat
  - The messages the governor sent
  - The messages other people sent to the governor
  - All metadata associated with a message, like the time of day a message was sent
- Anyone using this messaging app can create or edit content and produce a transcript that is indistinguishable from the original

<example fake screenshot (see Fig. 3)>

Tool-Tryable

- Politicians like Arden Fowler are required to use a messaging app that allows chat participants to edit anything in the chat
  - The messages the governor sent
  - The messages other people sent to the governor
  - All metadata associated with a message, like the time of day a message was sent
- Feel free to use the text message editing tool below to see how easy it is for someone to make anyone say anything

<interactive forger (see Fig. 4)>

Tool-Proactive

- Politicians like Arden Fowler are required to use a messaging app that allows chat participants to edit anything in the chat
  - The messages the governor sent
  - The messages other people sent to the governor
  - All metadata associated with a message, like the time of day a message was sent
- To prove that anyone can create text messages *supposedly* sent or received by Governor Fowler, the governor's IT department is constantly sending fake text messages among themselves

<example fake screenshot (see Fig. 3)>

Figure 2: The six different types of deniability evidence we used in our between-subjects (i.e., one condition per participant) study. Each condition lists a summary of information provided to participants, with references to <example fake screenshot> or <interactive forger> referring to images shown to participants or the interactive, in-survey screenshot generator, respectively.

3.3.2. Experts. In the next two conditions, denials are accompanied by statements from cryptographic experts about the deniability properties of a chat application and transcript. Specifically, the experts state that there is no cryptographic proof of authorship and anyone could have written the message (see Section 2.2). We test this notion in two separate conditions, one with and one without technical jargon.

In the EXPERT-JARGON condition, the defense offers expert opinion by two cryptographers from prestigious universities (MIT and Cambridge). The experts opine that the accusatory screenshot contains no "cryptographic proof" of authorship, thanks to the use of a "triple Diffie-Hellman handshake." Therefore, cryptographers conclude, anyone could have created a fake transcript (indistinguishable from a real transcript) and then taken a screenshot of it.

Given that the above statements contain jargon which is very likely to be foreign to a non-expert participant, we also wanted to assess whether explaining some of these technical terms could help with deniability; if users more fully understood what cryptographers were saying, would it improve deniability? To this end, the EXPERT-FRIENDLY condition mirrors EXPERT-JARGON, but uses less technical statements like "the app [used by the governor was] designed to leave no record" and "no record was left in this case."

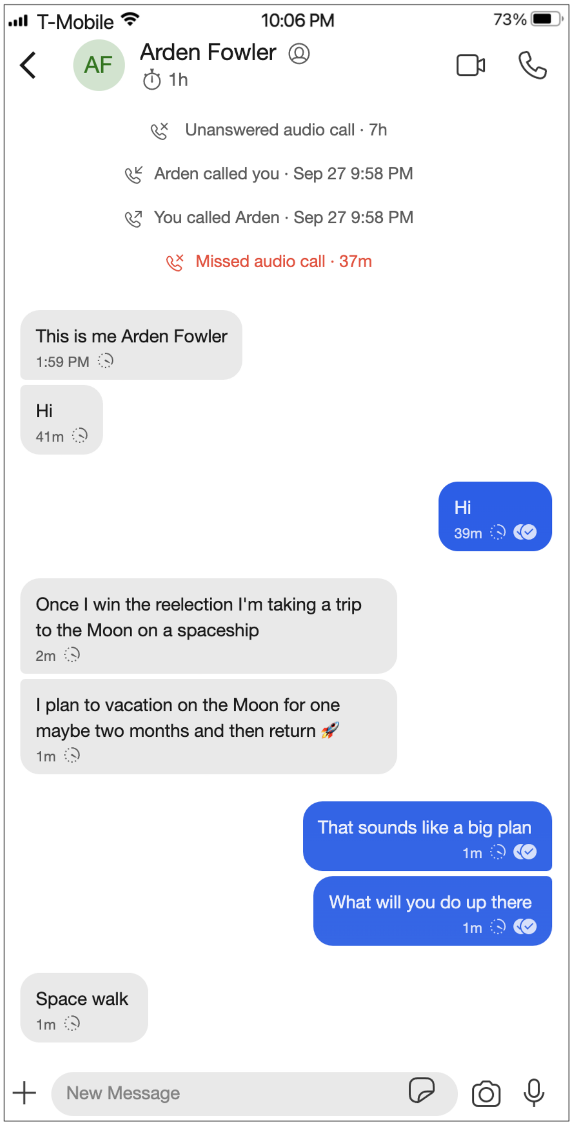
3.3.3. Tools. Our next three conditions explore the idea of making the theoretical possibility of message forgery more practical, by providing examples of existing forgeries.

These conditions describe a hypothetical "new" messaging app which allows chat participants to create or edit anything in a chat, including metadata such as call logs. Importantly, we also state that the governor is required to use this app, in order to insulate against the belief that use of this type of app is inherently suspicious [60]. This same hypothetical app is introduced in all of the tools-based conditions.

In the TOOL-EXISTS condition, we explain that the governor was using the "new" app, which makes screenshot forgery trivial. Participants are also given an example "fake" screenshot which looks identical to the one provided by the prosecution, but contains unrealistic information (the governor talks about vacationing on the moon, see Figure 3).

The TOOL-PROACTIVE condition is loosely inspired by the 2017 e-mail leaks associated with French President Emmanuel Macron [61]. Specifically, this condition looks at the case where an author proactively takes measures to aid deniability by themselves sending fake messages [62]. If an individual in a conversation thread mixes in "real" messages (i.e., validly sent and received messages which were intended to be truthful in content) with "fake" messages (i.e., validly sent and received messages which were intended to be untruthful in content) then a third party reading the messages post-facto can only guess at which messages are "real" and which are "fake." As such, a third party cannot rely on anything stated in a "leaked" message. Additionally, the fact that all messages were validly sent and received—and thus appear authentic even if nonsensical—can illustrate that just because a message "looks" real does not mean it is actually real.

Like the previous condition, participants are provided an example screenshot of a fake message supposedly authored proactively by the governor (the same one used in the TOOL-EXISTS condition, shown in Figure 3).



Unanswered audio call · 7h

Arden called you · Sep 27 9:58 PM

You called Arden · Sep 27 9:58 PM

Missed audio call · 37m

This is me Arden Fowler
1:59 PM

Hi
41m

Hi
39m

Once I win the reelection I'm taking a trip to the Moon on a spaceship
2m

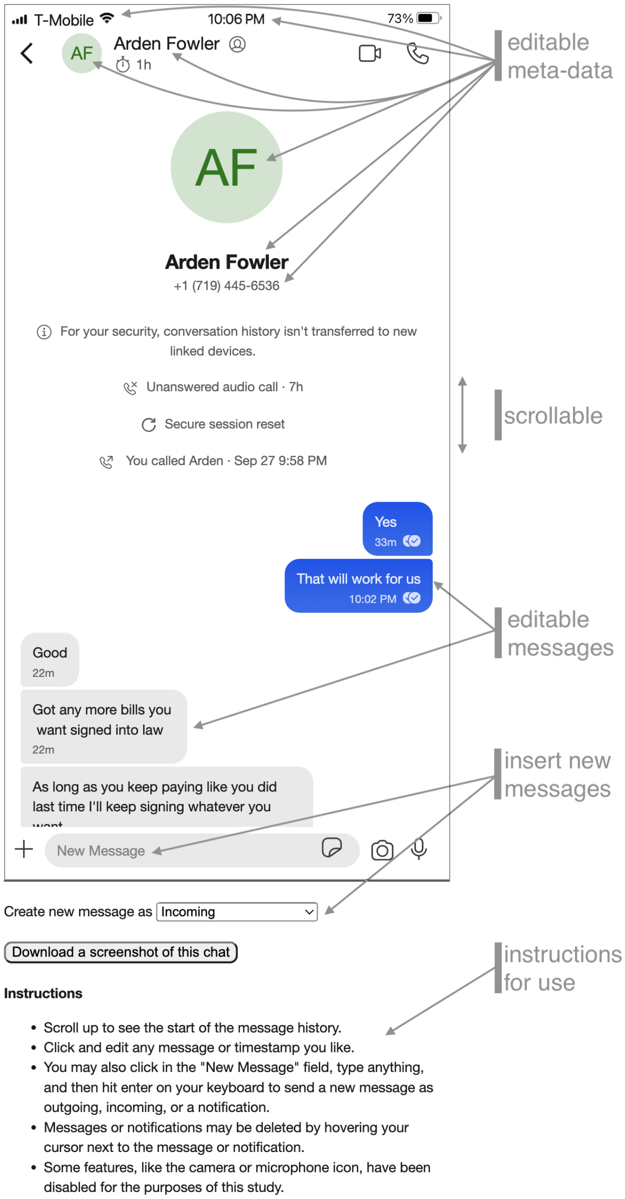
I plan to vacation on the Moon for one maybe two months and then return 🚀
1m

That sounds like a big plan
1m

What will you do up there
1m

Space walk
1m

Figure 3: Screenshot shown to participants as an example of a fake transcript that was easily generated by the forgery tool in the TOOL-EXISTS and TOOL-PROACTIVE conditions.



For your security, conversation history isn't transferred to new linked devices.

Unanswered audio call · 7h

Secure session reset

You called Arden · Sep 27 9:58 PM

Yes
33m

That will work for us
10:02 PM

Good
22m

Got any more bills you want signed into law
22m

As long as you keep paying like you did last time I'll keep signing whatever you want

Create new message as Incoming

Instructions

- Scroll up to see the start of the message history.
- Click and edit any message or timestamp you like.
- You may also click in the "New Message" field, type anything, and then hit enter on your keyboard to send a new message as outgoing, incoming, or a notification.
- Messages or notifications may be deleted by hovering your cursor next to the message or notification.
- Some features, like the camera or microphone icon, have been disabled for the purposes of this study.

Figure 4: Example of one type of deniability evidence provided to participants, allowing the editing of a screenshot within the survey itself (TOOL-TRYABLE). Participants were shown the interactive screenshot editor (without the right-aligned comments shown here), and could simply make these changes themselves.

The final condition, TOOL-TRYABLE, tests whether participants' ability to manipulate messages themselves impacts deniability. In this condition, participants, after being told about the "new" app used by the governor, were presented with an interactive screenshot generator within the survey itself (i.e., an embedded iFrame). The generator was based on a modified version of Signal Desktop, and was designed to allow editing, like the manipulation of text and the deletion or addition of messages [63]. An example of the sceenshot generator may be found in Figure 4. The UI was designed to look realistic, but sufficiently dissimilar from existing messaging applications to avoid bias.[5] We used the same screenshot generator tool to create all screenshots presented to participants.

3.4. Study recruitment

We used the Prolific crowdsourcing platform to recruit participants who were at least 18 years old, fluent in English, resided in the United States, and had at least a 95% platform approval rating. Because initial testing showed a severe gender imbalance, we used Prolific's gender-balancing filter

[5] The screenshot generation tool may be viewed and interacted with at https://github.com/nathanReitinger/deniability-GUI.

to achieve an approximately even split among men and women [64]. Participants were paid \$1.67 for completing the survey. **Survey 1** took, on average, 8.7 minutes; **Survey 2** took, on average, 8.2 minutes. Participants who failed both comprehension check questions (see Section 3.2), provided unreasonable free-text responses, or both were not paid and were discarded from the data.

University of Maryland's ethics review committee approved the study as "exempt" prior to recruitment. Participants were provided a consent form detailing study requirements, data retention procedures, and risks and benefits, and were able to withdraw at any time by simply discontinuing the survey. Participants were also debriefed as to the purpose of the study after completing the survey.

3.5. Data analysis

Here, we describe the statistical tests used to answer the research questions stated in Section 1. To understand whether the existence of any deniability evidence impacts participants' beliefs (RQ1) and final decisions (RQ3), as well as to understand the impact of personal and demographic factors such as gullibility or political leanings (RQ4), we use regression models.

We model the question of belief (five-point Likert scale, did Governor Fowler accept a bribe) using an ordinal logistic regression. For the participant's final decision (binary, should the governor be found guilty or not guilty), we use a logistic regression. For both regressions, we include the condition and five potential explanatory variables (shown in Table 1) as the input variables. To maximize explainability while also avoiding overfitting, we apply model selection by building models with all possible combinations of the input variables, always retaining the condition variable (as our primary variable of interest). We pick the model that minimizes the Akaike Information Criterion [65].

To compare all non-baseline conditions (RQ2), we use a two-tailed Mann-Whitney U test, appropriate for Likert data, for the question of belief [66]. For the final decision (RQ2, RQ4), we use Pearson's χ^2 test, appropriate for categorical data [67]. Holm correction for multiple comparisons was applied among all pairwise p-values [68].

We use qualitative coding to analyze free-text responses to the question: "How did the defense's evidence impact your belief about whether Governor Fowler took the bribe?" More specifically, we used inductive coding to create categories of "codes" which classified free-text responses [69]. One researcher made a codebook for this question, and a second coder then worked with the first to code 10% ($n = 60$) of the responses in tandem, revising the codebook and establishing a baseline. Both coders then independently coded batches of 10% of the dataset at a time, updating the codebook between each batch, until adequate reliability (defined as Cohen's $\kappa \geq 0.8$, or "almost perfect" [70]) was reached. After achieving sufficient reliability, the first coder coded all remaining responses, including re-coding any that had been analyzed with an earlier version of the codebook. To limit bias, free-text answers were coded without reference to the participant's experimental condition. After eight rounds of paired coding and codebook revision (480 responses), the coders reached a $\kappa = 0.86$. The resulting codebook may be found in Appendix B.

3.6. Limitations

Our study has several limitations common to human subjects research. First, we observed a tradeoff between providing more context for the bribery allegation—which would potentially be more realistic—and precisely testing only the deniability mechanisms of interest. This tradeoff between ecological validity and experimental precision is a standard limitation of survey experiments. We opted to present a minimal scenario, with little additional context, in order to maximize our ability to measure the effect of the deniability evidence itself. We do not attempt to measure how strong the effect of deniability evidence would be as compared to other kinds of evidence, pre-existing political beliefs, or other issues that mediate real-world judgment.

Second, our results might have been affected by several standard survey biases, including social desirability (e.g., attempting to respond in a pleasing way), satisficing (i.e., low-effort responses), and demand effects (i.e., responding with an inferred survey purpose in mind) [71]–[73]. We took steps to reduce these biases by including multiple comprehension-check questions, excluding low-quality responses pursuant to the exclusion criteria discussed in Section 3.4, and using a relatively low number of questions to reduce study fatigue. We were also careful to avoid any indication that the governor in our scenario was associated with any particular political party, state, or real-life politician. We piloted our survey to assess these initial concerns, making adjustments where necessary.

As with most crowdsourced samples, our sample is not fully representative of the U.S. population; we describe characteristics of our sample in detail in Section 4.1. Prior work has acknowledged this limitation but has also found crowdsourcing platforms to provide adequate sampling for security- and privacy-related topics [74]. Further, we limited our study to a U.S. context to limit political and cultural confounds; future work should consider the effectiveness of deniability evidence in other cultures and contexts. Nonetheless, we consider this a valuable first step toward understanding how cryptographic deniability affects perceptions in practice.

4. Results

In this section, we describe our sample, as well as the results of our initial and follow-up surveys.

4.1. Participants

As described in Section 3, our study consisted of two parts: **Survey 1**, in which we compared the six conditions introduced in Section 3.3, and **Survey 2**, in which we verified

Variable	Description	Possible values
Independent Variables		
Condition	One of the six experimental conditions, always retained.	Baseline: Baseline
Gullibility score	12-item scale (e.g., "[p]eople think I'm a little naive").	Range: 12–84 (continuous)
Political leaning	How would you describe your political views?	Buckets: moderate (baseline), liberal, conservative
Education	Highest level of education achieved.	Buckets: college degree (baseline), no college degree
Technical knowledge	Frequency of giving technology advice.	Buckets: often or more (baseline), sometimes or less
Age	How old are you?	Range: 18+ (continuous)
Dependent Variables		
Belief	Arden Fowler [did—did not] sign a bill in exchange for money.	Five points (Likert), definitely did to definitely did not
Decision	Arden Fowler [is—is not] guilty of accepting a bribe.	Two points (binary), is or is not

TABLE 1: Variables used in regression models. Model selection was used to pick among independent variables, with *condition* always retained. We select one final model per dependent variable.

the previous survey's results for a subset of the conditions with greater statistical power. In total, 635 individuals started **Survey 1**. Nine individuals were excluded due to failed attention checks, unreasonable free-text responses, or a combination of both; with dropout or timeout, 600 individuals, in total, successfully completed **Survey 1**. In **Survey 2**, 640 individuals started the survey and 600 completed it; eight participants were disqualified for failed attention checks, unreasonable free-text responses, or a combination of both, 32 individuals either dropped out or timed out.

	S1	S2		S1	S2
Gender			**Give Tech Advice**		
Female	50%	49%	Almost always	13%	14%
Male	48%	50%	Often	30%	25%
Other	3%	2%	Sometimes	41%	44%
Prefer not to say	1%	0%	Rarely	14%	14%
			Never	2%	3%
Ethnicity			**Education**		
White	76%	77%	Graduate/postgrad	13%	13%
Asian	12%	13%	College/undergrad	37%	40%
Hispanic or Latinx	11%	9%	Some college	24%	20%
Black or Af. Am.	8%	7%	Assoc. degree	10%	9%
Am. Ind. or AK Nat.	1%	1%	Vocational	2%	2%
Other	1%	1%	High sch. or equiv.	14%	15%
Prefer not to say	1%	1%	Some high sch.	1%	<1%
Political Affiliation			**Age**		
Very liberal	23%	20%	(18-30]	58%	48%
Liberal	36%	39%	(30-40]	22%	27%
Moderate	26%	27%	(65-100]	2%	2%
Conservative	12%	12%	(50-65]	8%	11%
Very conservative	3%	2%	(40-50]	11%	12%

TABLE 2: Participant demographics for **Survey 1** (*S1*) and **Survey 2** (*S2*), rounded to integers. Percentages may not add to 100% due to selection of multiple options by participants.

We summarize participant demographics in Table 2. As is common among crowdsourcing platforms, the participants were younger, whiter, more educated, and more technically savvy than the U.S. population as a whole [74]–[78]. A majority of participants in **Survey 1** (similar to **Survey 2**) also identified as liberal or very liberal (59.7%), compared to only 14.7% conservative or very conservative.[6]

Participants' average gullibility score (both surveys combined) was 31 of 84. The 25th, 50th, and 75th percentiles were 22, 29, and 37 for **Survey 1**, and 23, 29, 37 for **Survey 2**. While developing the scale, Teunisse et al. [59] found that scam victims scored around 41 and members of the "Skeptics Society" scored around 28.

We asked participants if they had ever faked a screenshot, or if not, whether they believed they could. While only 19.5% (both surveys combined) had faked a screenshot themselves, another 51.2% said they could if they wanted to, suggesting broad familiarity with the concept of fake images.

4.2. Effects of evidence on beliefs about deniability

After viewing the screenshot that served as evidence against the main character of our study, participants saw one of six different rebuttals that denied the prosecutor's accusations (Section 3.3). Immediately afterwards, we asked respondents whether they believed the allegations (i.e., whether the governor accepted the bribe). Participants answered on a five-point scale, indicating whether it was "definitely" or "probably" true, "definitely" or "probably" false, or whether the two outcomes were "equally likely."

Overall trends (RQ1) Results for the belief question are illustrated in Figure 5. In the Baseline condition ($n = 100$), where the deniability evidence was simply a claim that the accusation was false (with no evidence offered), only 5% said the governor probably did not take a bribe, and no participant responded that they were definite about this. In contrast, 66% said the governor definitely or probably did take a bribe. Less than a third, 29%, were undecided.

The picture was much different in all cases where evidence was presented (i.e., in every condition other than the Baseline). In all of these ($n = 100$ each), the plurality

[6] We considered trying to balance political leanings, but found this would reduce the pool of eligible participants significantly (from 53,219 to 3,172 when recruiting conservative-only participants); instead, political imbalance is a limitation of our sample (see Section 3.6).

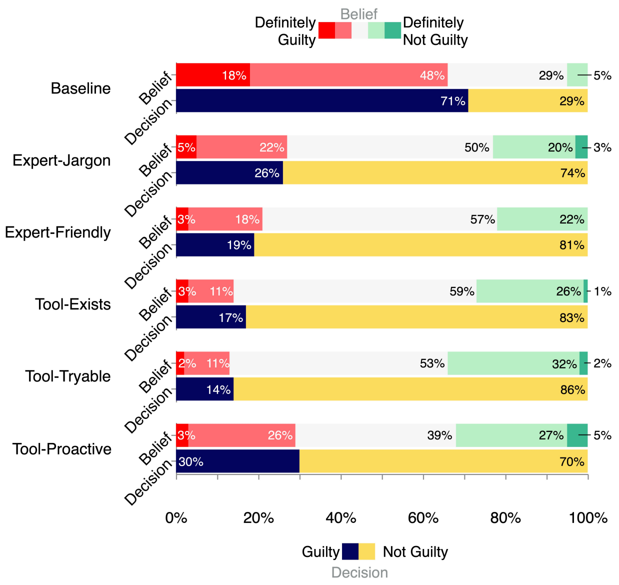


Figure 5: Fraction of respondents in **Survey 1** who believed the governor was guilty or not guilty (i.e., belief) stacked against fraction of respondents who acted on this belief in returning a verdict of guilty or not guilty (i.e., decision). Most participants outside of the BASELINE condition held the belief that it was equally likely the governor did or did not commit this crime (39-59%), and a majority of these participants ended up voting not guilty (70%+).

of participants (and as high as 59% in TOOL-EXISTS) stated that the claims for and against the governor were "equally likely." The fraction of respondents who believed the governor *did not* take the bribe was at least four times as high as in the BASELINE condition, with a minimum of 22% (in the EXPERT-FRIENDLY condition). In contrast, a maximum of 29% (in the TOOL-PROACTIVE condition) believed the evidence against the governor.

The condition that participants found most convincing was TOOL-TRYABLE (in which participants could themselves interactively generate fake screenshots, see Figure 4). This condition had the highest fraction of respondents who believed the governor was innocent (34%) and the lowest fraction of those believing the accusations (13%).

Effect size and other explanatory variables (RQ1, RQ4) We used an ordinal logistic regression to analyze people's beliefs about the evidence presented in our study (Table 3a). This served three purposes: it allowed us to verify the significance of the deniability evidence; it allowed us to estimate *how much* the evidence shifted beliefs; and it enabled us to test whether the effect could be explained by, or was correlated with, participant demographics or other personal characteristics. Table 1 lists the factors that were included in the initial regression model.

We found that the differences between BASELINE and every other evidence condition were statistically significant. Specifically, compared to BASELINE, participants in other conditions were 5.6–10.7 times as likely to increase one step on the five-point belief scale, toward believing the governor's denial. The TOOL-TRYABLE condition again emerged as the most persuasive (evidenced by the highest odds ratio), but not significantly more than the other deniability conditions (as seen in the overlapping confidence intervals).

	Odds Ratio	95% CI	p
Belief			
Condition (v. Baseline)			
EXPERT-JARGON	5.6	[3.3, 9.6]	**<0.001**
EXPERT-FRIENDLY	6.3	[3.7, 10.9]	**<0.001**
TOOL-EXISTS	8.5	[5.0, 14.7]	**<0.001**
TOOL-TRYABLE	10.7	[6.2, 18.6]	**<0.001**
TOOL-PROACTIVE	7.0	[4.0, 12.2]	**<0.001**
Demographic Covariates			
Gullibility	1.0	[1.0, 1.0]	0.122

(a) Ordinal logistic regression

	Odds Ratio	95% CI	p
Decision			
Condition (v. Baseline)			
EXPERT-JARGON	1.6	[1.4, 1.8]	**<0.001**
EXPERT-FRIENDLY	1.7	[1.5, 1.9]	**<0.001**
TOOL-EXISTS	1.7	[1.5, 1.9]	**<0.001**
TOOL-TRYABLE	1.8	[1.6, 2.0]	**<0.001**
TOOL-PROACTIVE	1.5	[1.3, 1.7]	**<0.001**
Demographic Covariates			
Politics (Conservative)	1.1	[1.0, 1.2]	0.084
Politics (Liberal)	1.0	[0.9, 1.1]	0.930

(b) Logistic regression

TABLE 3: Regression results from **Survey 1** for belief that the governor took a bribe (3a) and for decision (guilty or not guilty) (3b). For belief, the odds ratios show that participants in each non-baseline condition were 5.6–10.7× more likely to increase one step toward believing the governor's denial (compared to BASELINE), holding all other variables constant. For decision, participants in non-baseline conditions were 1.5–1.8× as likely to choose not guilty, compared to BASELINE. Coefficients were exponentiated to create Odds Ratios (OR); confidence intervals are $[2.5\%, 97.5\%]$; statistically significant p-values are noted in bold. Pseudo-R^2 (Aldrich-Nelson) for belief is 0.2 and for decision is 0.3.

None of the personal characteristics tested in our model (e.g., demographics, political views, or gullibility) were found to be significant.

Differences between types of evidence (RQ2) To investigate how the non-baseline conditions compared against each other and determine whether some types of evidence were more convincing than others, we followed up our regression with pairwise comparisons between all non-baseline conditions using Mann-Whitney U tests (Table 4a). While some differences were more pronounced than others, none were significant after applying Holm correction to account for comparing every pair of conditions. The same was true for our χ^2 test, none of the comparisons were significant

	EXPERT-JARGON	EXPERT-FRIENDLY	TOOL-EXISTS	TOOL-TRYABLE
EXPERT-FRIENDLY	1.000			
TOOL-EXISTS	0.794	1.000		
TOOL-TRYABLE	0.145	0.259	1.000	
TOOL-PROACTIVE	1.000	1.000	1.000	1.000

(a) MWU p-values of **Survey 1**

	EXPERT-JARGON	EXPERT-FRIENDLY	TOOL-EXISTS	TOOL-TRYABLE
EXPERT-FRIENDLY	—			
TOOL-EXISTS	—	—		
TOOL-TRYABLE	—	**<0.001**	0.276	
TOOL-PROACTIVE	—	—	—	—

(b) MWU p-values of **Survey 2**

TABLE 4: MWU results from **Survey 1** (Table 4a) and **Survey 2** (4b), showing pairwise comparisons among conditions. For **Survey 1**, all non-BASELINE conditions were compared to each other, and the table shows p-values after Holm correction was applied. For **Survey 2**, only two comparisons were made and no correction was required. Statistically significant p-values noted with bold.

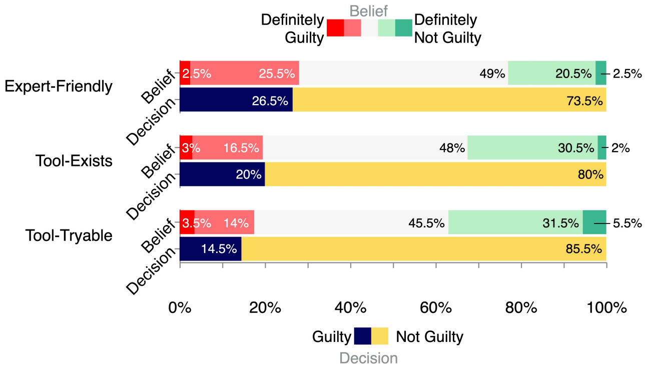


Figure 6: Fraction of respondents in **Survey 2** who believed the governor was guilty or not guilty (i.e., belief), and who returned a verdict of guilty or not guilty (i.e., decision).

after correction. Because we did observe some plausibly meaningful trends, we decided to pursue this question in more depth by conducting a follow-up study; its results are reported next in Section 4.3.

4.3. Replication study (RQ2)

As described above, when we compared the different evidence conditions to each other, we observed some apparent trends, but found no statistical significance after correcting for multiple comparisons. We therefore decided to perform a follow-up study ($n = 600$) to examine these effects in more detail. The most likely conditions to show significance (see Table 4 (**Survey 1**)) were the comparisons between ⟨TOOL-EXISTS, EXPERT-JARGON⟩ and ⟨TOOL-TRYABLE, EXPERT-JARGON⟩ and ⟨TOOL-TRYABLE, EXPERT-FRIENDLY⟩.

In **Survey 2**, we focused on three of these four conditions and increased statistical power (by doubling participant counts to 200 per condition). Specifically, we planned to compare TOOL-EXISTS to each of EXPERT-FRIENDLY and TOOL-TRYABLE. Given only two planned comparisons, we did not need to correct p-values [79], [80].

The beliefs and decisions exhibited in the replication study were quite similar to those in **Survey 1** (see Figure 6). All deniability evidence pushed at least 20% of participants to believe in the governor's innocence and a large majority to vote "not guilty." Once again, the tool-based deniability evidence was more convincing than testimony from experts. Also as before, a greater number of people in the TOOL-TRYABLE condition were convinced by the evidence than in TOOL-EXISTS, but only by a slight margin.

Using Mann-Whitney U tests (Table 4b), we found that the difference between EXPERT-FRIENDLY and TOOL-TRYABLE was significant ($\alpha = 0.05$), but the difference between TOOL-TRYABLE and TOOL-EXISTS was not.

4.4. Willingness to act on beliefs (RQ3)

One of our research goals was to study how deniability impacted beliefs versus (intended) decisions. In other words, are people willing to make decisions based on the doubt created in their minds by evidence of deniability? To study this, in addition to asking participants about whether they believed the evidence presented in our scenario, we asked about how they would vote if they were a jury member in this trial: guilty or not guilty. As described in Section 3.1, a verdict is a somewhat unique type of decision, but nonetheless asking this question gives us some insight into the gap between belief and action.

The results may be seen in Figure 5. Across all conditions, approximately everyone who thought the governor "definitely" or "probably" took a bribe opted for a "guilty" verdict. As expected, those who believed the counter-evidence generally voted "not guilty." Interestingly, nearly everyone who stated that the two outcomes were "equally likely" also voted "not guilty." It appears, therefore, that voting to convict required greater certainty than voting to acquit. This may reflect the general social-psychology understanding that decisions are harder to affect than beliefs [55]–[57], but it also likely reflects the seriousness of our courtroom scenario, in which voting to convict has life-changing consequences for our imaginary governor.

We modeled people's answers to this question using a similar regression as for the beliefs question (Table 3b). Overall, we found similar, but weaker, effects when looking at verdicts versus beliefs. In this case, compared to the baseline condition, participants were 1.5–1.8 times as likely to choose not guilty. Here as well demographics and other factors were not significant.

We also used χ^2 tests to compare non-baseline conditions. Although similar trends held in absolute numbers (tools more effective than experts), we found no statistically significant differences in either **Survey 1** or **Survey 2**.

4.5. Participant opinions

In addition to collecting people's beliefs and decisions in the hypothetical trial scenario, we directly asked participants: how did the defense's evidence impact your belief? We used qualitative coding to categorize their answers, which add context to our primarily quantitative results, but should not be interpreted as standalone results [81].

A summary of our results may be found in Table 5. We make several observations, all of which mirror our statistical results. A majority of participants in the BASELINE condition felt that the defense's evidence had *no impact* on their opinion, and no one felt it had a large impact.

Interestingly, 10% of participants in the BASELINE condition—much higher than in other conditions—indicated that the defense had the opposite of the intended effect and made them more likely to believe the accusation. This was explained by some as the lack of evidence making the accusation more convincing (e.g., P-23: "The defense's lack of evidence presented made me think the text was real.").

In line with our statistical results (see Section 4.3), we see that across all non-BASELINE conditions, the vast majority of participants said the evidence had either a small or a large effect (in the direction of believing the governor's denial). More participants in the TOOL-TRYABLE and TOOL-EXISTS conditions said the defense's evidence had a large effect on their decision or even made them change their mind, when compared to the other conditions. Further, there was minimal difference between TOOL-TRYABLE and TOOL-EXISTS. Overall, most responses in the non-BASELINE conditions reported a small effect.

We also asked participants what *other* evidence would have changed their minds about the governor's guilt. We hypothesized that this question would give us insight into the types of deniability evidence that might be most convincing to users. Analyzing these comments informally, we found that most participants looked for evidence that would be difficult or impossible to provide through a secure messaging app. For instance, one participant would change their mind based on "irrefutable evidence in the form of money trails, sworn testimony of witnessing parties [or] prior history of questionable practices" (P-7), sentiments which were expressed in piecemeal form by many other participants. Likewise, another participant would look to "voice recordings, documents, [and] phone calls" (P-78).

5. Discussion

Overall, we make two primary findings. First, the ability to deny having made a statement is improved when either experts confirm that there is no cryptographic proof of authorship or a tool which makes the creation of forged screenshots practical is involved. Second, these two types of evidence are not interchangeable; a screenshot-forgery tool provides a statistically significantly more effective denial than an expert's statement.

Deniability is not accepted by default Today's society seems primed for skepticism: "photoshopping" has entered the dictionary [82], knowledge of deepfakes is becoming more common [20]–[23], and the majority of Americans believe "fake news" is a problem [20]. In this environment, the cryptographic model of deniability seems like it might be universal: absent cryptographic evidence, all accusations are false. Our study strongly suggests otherwise. When presented (in a courtroom scenario) with a simple denial regarding authorship of a screenshot, a vast majority of participants believed the screenshot over the denial—despite there being no cryptographic proof (or any other evidence) that the individual actually wrote the message. In other words, the landscape of what deniability means to humans does not necessarily match how cryptographers might think about it.

Deniability requires evidence When evidence exceeds a bare assertion, deniability is more effective: our participants became more likely to believe a denial. In fact, participants in all conditions except BASELINE were five to ten times more likely to report increased belief in the denial—with TOOL-TRYABLE showing the most promising results. These findings suggest deniability is achievable and, therefore, a worthwhile objective. We further observe that different types of deniability evidence can have different effects.

Protocol deniability is largely convincing We wanted to test whether the current way cryptographers think of deniability—the absence of a cryptographic proof that the sender authored a message—could be convincing. We encapsulated this in two expert-based conditions. We found that people are receptive to this argument, with nearly three quarters of participants finding this evidence at least partially convincing. This result suggests that existing secure messaging protocols, which incorporate deniability, may be able to achieve their aim of convincing people that messages in any given conversation may have been forged.

Clarifying jargon does not enhance deniability In our jargon-filled expert condition (EXPERT-JARGON), the evidence presented to participants included cryptographic terms such as "triple Diffie-Hellman handshake," which most non-experts are unlikely to be familiar with. We wanted to test whether using more comprehensible terminology, which explained the same notions in an easier-to-understand way, would make people more receptive to the deniability argument. Surprisingly, we found that participants trusted the words of experts regardless of how comprehensible the experts' statements were. The beliefs we observed in the more comprehensible condition (EXPERT-FRIENDLY) were hardly any different from those in the less comprehensible condition (EXPERT-JARGON). This suggests that deniability is not necessarily more attainable when comprehension is improved.

A forgery tool is more convincing than statements about protocols We found that referencing tools which allow chat participants to edit message transcripts improves deniability more than experts making statements about how a

		% of participants					
Did evidence impact belief?	**Representative quote**	**BASELINE**	**EXPERT-JARGON**	**EXPERT-FRIENDLY**	**TOOL-EXISTS**	**TOOL-TRYABLE**	**TOOL-PROACTIVE**
Opposite of intended effect	[TOOL-PROACTIVE] [I]t actually showed it was more likely he wrote it, his staff would never do that as a test. (P-39)	10%	1%	1%	0%	0%	3%
No effect	[BASELINE] [H]e did not provide any evidence to support his claim, so it did not impact my belief in any way. (P-33)	45%	6%	7%	9%	8%	3%
Ambiguous (either no effect or small effect)	[EXPERT-JARGON] It didn't impact it much. While they say there is no cryptographic evidence of who the author is, they don't discount that the author could still be Fowler. (P-44)	16%	5%	3%	6%	1%	2%
Small effect	[TOOL-EXISTS] It made me think that it was more likely that the first screenshot was faked and he may not have taken the bribe. (P-75)	24%	70%	77%	60%	61%	72%
Large effect	[TOOL-EXISTS] It seems very real and it is clearly seen that other types of conversations can be created to make others seem guilty, I think it is risky to use an application if you can edit the messages, it is very likely that many false accusations can be made, not only to Arden Fowler. (P-45)	0%	11%	11%	23%	23%	14%
Changed belief (from guilty to either equally likely or not guilty)	[TOOL-TRYABLE] It changed my mind that Gov. Fowler was definitely guilty to a 50/50% change of guilt. If [I] used the messaging app and saw it was easy to fake a text then I would have serious doubt to the validity of the text message. (P-86)	1%	2%	1%	2%	4%	1%
Other	[TOOL-TRYABLE] Why in the world would "politicians like Arden Fowler [be] required to use a new, special messaging app that helps neutralize disinformation campaigns." Just to give themselves an out to anything? (P-54)	4%	5%	0%	0%	3%	5%

TABLE 5: Percentage of participants per condition describing how the deniability evidence impacted their belief about whether Governor Fowler took the bribe. Appendix B provides more detail on the codes and how they were generated.

protocol works. This trend from **Survey 1** was confirmed in **Survey 2**, showing a statistically significant difference between EXPERT-FRIENDLY and TOOL-TRYABLE. Several participants expressed sentiments similar to P-92: "Just because there is no cryptographic proof does not mean that it has been faked." This suggests that the current way deniability is thought of by cryptographers may be necessary, but not maximally effective, for convincing non-experts: the success of the denial improves when the possibility of a forgery is practical, not just theoretical.

First-hand forgery experience does not make a big difference We hypothesized that being able to use an interactive tool to produce fake transcripts would increase an individual's confidence in the possibility of forgeries by tangibly demonstrating the simplicity of producing them. We tested this by providing such a tool in the TOOL-TRYABLE condition. To our surprise, we found that this first-hand experience did not make a significant difference in comparison to other conditions: only slightly more people were convinced by it when compared with the TOOL-EXISTS condition, where a single faked screenshot was entered into evidence along with a statement about the existence of the forgery tool. It appears that what really matters to people is demonstrating that editing a transcript is not just a theoretical possibility, but that it is practical and easily accomplished. Relative to this knowledge that forgeries are easy, first-hand experience seems to be less important.

Proactive effort is effective, but suspicious We also tested a variant of the forgery tool that described fake messages being created proactively while the app was being used (TOOL-PROACTIVE). The scenario was meant to emphasize that real and fake messages were impossible to distinguish from the very start. This type of deniability evidence produced polarizing and somewhat contradictory responses: it had the most "definitely not guilty" responses, but conversely, this condition was also second for the most "definitely guilty" responses (excluding the baseline, see Figure 5). There are several possible explanations. One is that use of a proactive forgery tool is inherently suspicious, despite our statements that the tool's use was required.

Another possibility, suggested in some open-ended responses, is that our intentionally silly example of a proactively forged message (i.e., the message content referenced moonwalking) made it harder to take the notion of forgeries seriously. That is, some participants seemed to assume that if all proactive forgeries were silly, differentiating real and fake messages would be straightforward. Further it may be difficult to believe that the governor's staff would purposefully forge a message as incriminating as our bribery example, even "as a test" (P-39). Such comments suggest that participants in the TOOL-PROACTIVE condition absorbed the idea of proactive forgeries, but did not necessarily make the further inference that if the governor's staff can forge messages, others can too.

Our results suggest building automatically-created, proactive forgeries into a messaging app has promise, but what kind of messages to forge and how to generalize from author to third-party forgeries are important considerations.

Open questions If secure messaging providers like Signal want to offer users the best possible chance at deniability, then we need to look beyond deniability as a "lack of proof of authorship." Our paper shows that tools-based deniability evidence hold promise, but open questions remain. Would our results hold in a non-courtroom, interpersonal setting? If plausible deniability is attainable, do its advantages outweigh its potential for abuse? Is there a point when too much ambiguity becomes detrimental to a deniable protocol (i.e., the trustworthiness–deniability trade-off)? We leave these questions for future work.

Acknowledgments

We thank our participants for contributing data to this research, as well as the reviewers for their helpful comments. This work received partial support from a Facebook Research award.

References

[1] WikiLeaks. The Podesta emails. https://web.archive.org/web/20210307231437/https://wikileaks.org/podesta-emails/?q=&mfrom=&mto=&title=&notitle=&date_from=&date_to=&nofrom=&noto=&count=50&sort=6&page=2&#searchresult.

[2] AtoZ Wiki. Timeline of the 2012 United States presidential election. https://atozwiki.com/Timeline_of_the_2012_United_States_presidential_election.

[3] Mike Masnick, "The Clinton campaign should stop denying that the WikiLeaks emails are valid; they are and they're real," https://www.techdirt.com/2016/10/25/clinton-campaign-should-stop-denying-that-wikileaks-emails-are-valid-they-are-theyre-real/, TechDirt, 2016.

[4] BBC, "18 revelations from WikiLeaks' hacked Clinton emails," https://www.bbc.com/news/world-us-canada-37639370, BBC, 2016.

[5] Lauren Carroll, "Are the Clinton WikiLeaks emails doctored, or are they authentic?" https://www.politifact.com/article/2016/oct/23/are-clinton-wikileaks-emails-doctored-or-are-they-/, PolitiFact, 2016.

[6] Barry Leiba and Jim Fenton, "DomainKeys Identified Mail (DKIM): Using digital signatures for domain verification." In *Proc. CEAS*, 2007.

[7] 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/.

[8] Estelle Derouet, "Fighting phishing and securing data with email authentication," *Computer Fraud & Security*, vol. 2016, no. 10, pp. 5–8, 2016.

[9] WikiLeaks. https://wikileaks.org/DKIM-Verification.html.

[10] Craig Timberg, Matt Viser, and Tom Hamburger, "Here's how The Post analyzed Hunter Biden's laptop," https://www.washingtonpost.com/technology/2022/03/30/hunter-biden-laptop-data-examined/, The Washington Post, 2022.

[11] Nikita Borisov, Ian Goldberg, and Eric Brewer, "Off-The-Record communication, or, why not to use PGP." In *Proc. WPES*, 2004.

[12] Jon Callas, Lutz Donnerhacke, Hal Finney, and Rodney Thayer. OpenPGP message format. https://datatracker.ietf.org/doc/html/rfc2440.

[13] Blake Ramsdell. S/MIME version 3 message specification. https://www.rfc-editor.org/rfc/rfc2633.

[14] Jon Callas, Lutz Donnerhacke, Hal Finney, David Shaw, and Rodney Thayer. OpenPGP message format. https://www.rfc-editor.org/rfc/rfc4880.

[15] moxie0. Simplifying OTR deniability. https://signal.org/blog/simplifying-otr-deniability/.

[16] Nik Unger, Sergej Dechand, Joseph Bonneau, Sascha Fahl, Henning Perl, Ian Goldberg, and Matthew Smith, "Sok: Secure messaging." In *Proc. IEEE S&P*, 2015.

[17] Tole Sutikno, Lina Handayani, Deris Stiawan, Munawar Agus Riyadi, and Imam Much Ibnu Subroto, "WhatsApp, Viber and Telegram: Which is the best for instant messaging?" *International Journal of Electrical & Computer Engineering*, vol. 6, no. 3, pp. 909–914, 2016.

[18] Nihal Vatandas, Rosario Gennaro, Bertrand Ithurburn, and Hugo Krawczyk, "On the cryptographic deniability of the Signal protocol." In *Proc. IACR*, 2020.

[19] Manish Singh, "Signal's Brian Acton talks about exploding growth, monetization, and WhatsApp data sharing outrage," https://techcrunch.com/2021/01/12/signal-brian-acton-talks-about-exploding-growth-monetization-and-whatsapp-data-sharing-outrage/, TechCrunch, 2021.

[20] Michael Barthel, Amy Mitchell, and Jesse Holcomb, "Many Americans believe fake news is sowing confusion," https://www.pewresearch.org/journalism/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/, Pew Research Center, 2016.

[21] Justin D Cochran and Stuart A Napshin, "Deepfakes: Awareness, concerns, and platform accountability," *Cyberpsychology, Behavior, and Social Networking*, vol. 24, no. 3, pp. 164–172, 2021.

[22] Mika Westerlund, "The emergence of deepfake technology: A review," *Technology Innovation Management Review*, vol. 9, no. 11, pp. 39–52, 2019.

[23] Rashid Tahir, Brishna Batool, Hira Jamshed, Mahnoor Jameel, Mubashir Anwar, Faizan Ahmed, and Muhammad Adeel Zaffar, "Seeing is believing: Exploring perceptual differences in deepfake videos." In *Proc. CHI*, 2021.

[24] Elizabeth Fricker, *Against gullibility*. Kluwer Academic Publishers, 1994.

[25] Misha Müller. Plausible deniability: From Gricean pragmatics to the insights of relevance theory. https://www.academia.edu/29197017/Plausible_Deniability_From_gricean_pragmatics_to_the_insights_of_Relevance_theory?source=swp_share.

[26] Steven Pinker, Martin A. Nowak, and James J. Lee, "The logic of indirect speech." In *PNAS*, 2008.

[27] Andrew Peet, "Testimony, pragmatics, and plausible deniability," *Episteme*, vol. 12, no. 1, pp. 29–51, 2015.

[28] Kenneth Church and Ramesh Patil, “Coping with syntactic ambiguity or how to put the block in the box on the table,” *American Journal of Computational Linguistics*, vol. 8, no. 3-4, pp. 139–149, 1982.

[29] Rory Cormac and Richard J Aldrich, “Grey is the new black: Covert action and implausible deniability,” *International Affairs*, vol. 94, no. 3, pp. 477–494, 2018.

[30] Document 292: National Security Council directive on Office of Special Projects. https://history.state.gov/historicaldocuments/frus1945-50Intel/d292.

[31] Gregory F. Treverton, “Covert action and open society,” *Foreign Affairs*, vol. 65, no. 5, pp. 995–1014, 1986.

[32] Michael Poznansky, “Revisiting plausible deniability,” *Journal of Strategic Studies*, vol. 45, no. 4, pp. 511–533, 2020.

[33] *Report of the congressional committees investigating the Iran-Contra affair*. United States Government Publishing Office, 1987.

[34] Rein Canetti, Cynthia Dwork, Moni Naor, and Rafail Ostrovsky, “Deniable encryption.” In *Proc. CRYPTO*, 1997.

[35] Nik Unger and Ian Goldberg, “Deniable key exchanges for secure messaging.” In *Proc. CCS*, 2015.

[36] Nik Unger and Ian Goldberg, “Improved strongly deniable authenticated key exchanges for secure messaging.” In *Proc. PETS*, 2018.

[37] Ksenia Ermoshina and Francesca Musiani, “‘Standardising by running code’: The Signal protocol and de facto standardisation in end-to-end encrypted messaging,” *Internet Histories*, vol. 3, no. 3-4, pp. 343–363, 2019.

[38] Moxie Marlinspike and Trevor Perrin, “The x3dh key agreement protocol,” https://signal.org/docs/specifications/x3dh/x3dh.pdf, Signal, 2016.

[39] Michael A Specter, Sunoo Park, and Matthew Green, “KeyForge: Non-attributable email from forward-forgeable signatures.” In *Proc. USENIX Security*, 2021.

[40] Gabrielle Beck, Arka Rai Choudhuri, Matthew Green, Abhishek Jain, and Pratyush Ranjan Tiwari, “Time-deniable signatures,” https://eprint.iacr.org/2022/1018, Cryptology ePrint Archive, 2022.

[41] Alma Whitten and Doug J Tygar, “Why Johnny can’t encrypt: A usability evaluation of PGP 5.0.” In *Proc. USENIX Security*, 1999.

[42] 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 *Proc. EuroUSEC*, 2016.

[43] Joshua Tan, Lujo Bauer, Joseph Bonneau, Lorrie Faith Cranor, Jeremy Thomas, and Blase Ur, “Can unicorns help users compare crypto key fingerprints?” In *Proc. CHI*, 2017.

[44] Amir Herzberg, Hemi Leibowitz, Kent Seamons, Elham Vaziripour, Justin Wu, and Daniel Zappala, “Secure messaging authentication ceremonies are broken.” In *Proc. IEEE S&P*, 2020.

[45] Elham Vaziripour, Justin Wu, Mark O’Neill, Ray Clinton, Jordan Whitehead, Scott Heidbrink, Kent Seamons, and Daniel Zappala, “Is that you, Alice? A usability study of the authentication ceremony of secure messaging applications.” In *Proc. SOUPS*, 2017.

[46] Amir Herzberg and Hemi Leibowitz, “Can Johnny finally encrypt? Evaluating E2E-encryption in popular IM applications.” In *Proc. STAST*, 2016.

[47] Francesca Musiani and Ksenia Ermoshina, “What is a good secure messaging tool? The EFF secure messaging scorecard and the shaping of digital (usable) security,” *Westminster Papers in Communication and Culture*, vol. 12, no. 3, pp. 51–71, 2017.

[48] Daniel V Bailey, Philipp Markert, and Adam J Aviv, “‘I have no idea what they’re trying to accomplish’: Enthusiastic and casual Signal users’ understanding of Signal PINs.” In *Proc. SOUPS*, 2021.

[49] Christine Geeng, and Jevan Hutson, and Franziska Roesner, “Usable Sexurity: Studying people’s concerns and strategies when sexting.” In *Proc. SOUPS*, 2020.

[50] Sascha Fahl, Marian Harbach, Thomas Muders. Matthew Smith, and Uwe Sander, “Helping Johnny 2.0 to encrypt his Facebook conversations.” In *Proc. SOUPS*, 2012.

[51] Ksenia Ermoshina, and Harry Halpin, and Francesca Musiani, “Can Johnny build a protocol? Co-ordinating developer and user intentions for privacy-enhanced secure messaging protocols.” In *Proc. EuroUSEC*, 2017.

[52] Luke Rosiak, “Here’s cryptographic proof that Donna Brazile is wrong, WikiLeaks emails are real,” https://dailycaller.com/2016/10/21/heres-cryptographic-proof-that-donna-brazile-is-wrong-wikileaks-emails-are-real/, Daily Caller, 2016.

[53] Jose Pagliery and Roger Sollenberger, “Bombshell letter: Gaetz paid for sex with minor, Wingman says,” https://www.thedailybeast.com/joel-greenberg-letter-written-for-roger-stone-says-matt-gaetz-paid-for-sex-with-minor, Daily Beast, 2021.

[54] Ralph Artigliere, “Sequestration for the twenty-first century: Disconnecting jurors from the Internet during trial,” *Drake Law Review*, vol. 59, pp. 621–647, 2010.

[55] Timote M Vaioleti, “Talanoa research methodology: A developing position on Pacific research,” *Waikato Journal of Education*, vol. 12, pp. 21–34, 2006.

[56] Allen E Liska, “A critical examination of the causal structure of the Fishbein/Ajzen attitude–behavior model,” *Social Psychology Quarterly*, vol. 47, no. 1, pp. 61–74, 1984.

[57] Tanzhe Tang and Caspar G Chorus, “Learning opinions by observing actions: Simulation of opinion dynamics using an action-opinion inference model,” *Journal of Artificial Societies and Social Simulation*, vol. 22, no. 3, 2019.

[58] 18 U.S.C. § 201(b)(2) (2012).

[59] Alessandra K Teunisse, Trevor I Case, Julie Fitness, and Naomi Sweller, “I Should have known better: Development of a self-report measure of gullibility,” *Personality and Social Psychology Bulletin*, vol. 46, no. 3, pp. 408–423, 2020.

[60] Shirley Gaw, Edward W Felten, and Patricia Fernandez-Kelly, “Secrecy, flagging, and paranoia: Adoption criteria in encrypted email.” In *Proc. CHI*, 2006.

[61] Jean-Baptiste Jeangène Vilmer, “The ‘#Macron leaks’ operation: A post-mortem,” https://www.atlanticcouncil.org/in-depth-research-reports/report/the-macron-leaks-operation-a-post-mortem/, Atlantic Council, 2019.

[62] Adam Nossiter, David E Sanger, and Nicole Perlroth, “Hackers came, but the French were prepared,” https://www.nytimes.com/2017/05/09/world/europe/hackers-came-but-the-french-were-prepared.html, The New York Times, 2017.

[63] Signal: A private messenger for Windows, macOS, and Linux. https://github.com/signalapp/Signal-Desktop.

[64] Nick Charalambides, “We recently went viral on TikTok—here’s what we learned,” https://blog.prolific.co/we-recently-went-viral-on-tiktok-heres-what-we-learned/, Prolific, 2021.

[65] Hamparsum Bozdogan, “Model selection and Akaike’s information criterion (AIC): The general theory and its analytical extensions,” *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.

[66] Patrick E McKnight and Julius Najab, *Mann–Whitney U Test*. Wiley Online Library, 2010.

[67] Karl Pearson, “On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

[68] Sture Holm, “A simple sequentially rejective multiple test procedure,” *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[69] Johnny Saldana, *The coding manual for qualitative researchers*. SAGE Publications, 2021.

[70] Richard J Landis, and Gary G Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol. 33, pp. 159–174, 1977.

[71] Jon A Krosnick, Sowmya Narayan, and Wendy R Smith, “Satisficing in surveys: Initial evidence,” *New Directions for Evaluation*, vol. 1996, no. 70, pp. 29–44, 1996.

[72] Ivar Krumpal, “Determinants of social desirability bias in sensitive surveys: A literature review,” *Quality & Quantity*, vol. 47, no. 4, pp. 2025–2047, 2013.

[73] Jonathan Mummolo and Erik Peterson, “Demand effects in survey experiments: An empirical assessment,” *American Political Science Review*, vol. 113, no. 2, pp. 517–529, 2019.

[74] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek, “How well do my results generalize? Comparing security and privacy survey results from Mturk, web, and telephone samples.” In *Proc. IEEE S&P*, 2019.

[75] Djellel Difallah, Elena Filatova, and Panos Ipeirotis, “Demographics and dynamics of Mechanical Turk workers.” In *Proc. WSDM*, 2018.

[76] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson, “Who are the Turkers? Worker demographics in Amazon Mechanical Turk.” In *Proc. CHI EA*, 2010.

[77] Anne M. Turner, Thomas Engelsma, Jean O. Taylor, Rashmi K. Sharma, and George Demiris, “Recruiting older adult participants through crowdsourcing platforms: Mechanical Turk versus Prolific Academic.” In *Proc. AMIA*, 2020.

[78] Jenny Tang, Eleanor Birrell, and Ada Lerner, “Replication: How well do my results generalize now? The external validity of online privacy and security surveys.” In *Proc. SOUPS*, 2022.

[79] Ronald J Feise, “Do multiple outcome measures require p-value adjustment?” *BMC Medical Research Methodology*, vol. 2, no. 8, pp. 1–4, 2002.

[80] Mohieddin Jafari and Naser Ansari-Pour, “Why, when and how to adjust your P values?” *Cell Journal*, vol. 20, no. 4, pp. 604–607, 2019.

[81] Carolyn B Seaman, “Qualitative methods in empirical studies of software engineering.” In *Proc. IEEE TSE*, 1999.

[82] Merriam Webster. Photoshop. https://www.merriam-webster.com/dictionary/photoshop.

Appendix A. Survey

[`consent and introduction`]

Imagine the following scenario:

You were recently selected to serve on a jury. As a jury member, you are instructed to be impartial when making decisions. This means that you should not rely on external sources of information like personal bias or preconceived beliefs about individuals or situations; you must base your verdict only on the evidence presented.

The trial involves the current Governor of your state, Arden Fowler. The trial centers around a piece of evidence referred to as “The Leak”—a leaked text message screenshot claimed to be authored by Arden Fowler.

The prosecution claims that the Governor made the statements found in the screenshot, and therefore committed bribery. The text message screenshot, as provided by the prosecution, has been reproduced below:

[`Screenshot (see Figure 1)`]

What do you think this screenshot is about?

□ The right of citizens to bear arms
□ Governor Fowler signing a bill
□ Governor Fowler’s opinion on Taylor Swift
□ Governor Fowler paying utility bills online

1—Baseline

The defense argues that Arden Fowler did not make these statements, and therefore did not commit bribery. The defense states:

This text message screenshot is fake.

This block (R1-R3) repeats for each condition

<R1> **The defense’s argument could be summarized as:**

□ The screenshot is fake because it was sunny outside on October 8, especially midday around 11 AM or noon, when these messages were sent
□ The screenshot is fake because the Governor doesn’t use T-Mobile
□ The screenshot is real because the Governor is on trial
□ The screenshot is fake because the defense says it is fake

<R2 (belief)> **Which do you agree most with:**

□ Arden Fowler definitely agreed to sign a bill in exchange for money
□ Arden Fowler probably agreed to sign a bill in exchange for money
□ I think it’s equally likely that Arden Fowler did or did not sign a bill in exchange for money
□ Arden Fowler probably did not agree to sign a bill in exchange for money
□ Arden Fowler definitely did not agree to sign a bill in exchange for money

<R3> **How did the defense’s evidence (Exhibit B) impact your belief about whether Governor Fowler took the bribe?** [free-text]

2—Expert-Jargon

The defense argues that Arden Fowler did not make these statements, and therefore did not commit bribery. The defense states:

This text message screenshot is fake.

Out of an abundance of caution, we had two experts in cryptography (award-winning cryptography professors from MIT and the University of Cambridge) look into the message and its associated information. The experts both agreed that the leaked screenshot contains “no cryptographic proof” of authorship.

The experts certified that the messaging app which was used to create this message relied on a “triple Diffie-Hellman handshake,” which leaves no cryptographic evidence of who might have authored a particular message after the fact.

Cryptographically, anyone can fake a transcript by merely knowing the public keys of anyone else, which are publicly available. No one can tell the difference between a real transcript and an edited transcript.

As a result, anyone could have created a fake message like this and then taken a screenshot of it.

[R1–R3*]

□ *(R1-d) The screenshot is fake because the experts say there is “no cryptographic proof” of authorship

---

3—EXPERT-FRIENDLY

The defense argues that Arden Fowler did not make these statements, and therefore did not commit bribery. The defense states:

This text message screenshot is fake.

Out of an abundance of caution, we had two experts in cryptography (award-winning cryptography professors from MIT and Cambridge) look into the message and its associated information. The experts certified that the messaging app which was used to create this message is designed to leave no record, and no record was left in this case. Anyone can fake a transcript and create a new transcript with edited content that is indistinguishable from the original transcript.

As a result, anyone could have created a fake message like this and then taken a screenshot of it.

[R1–R3*]

□ *(R1-d) The screenshot is fake because the experts say anyone can create fake messages that are identical to real messages

---

4—TOOL-EXISTS

The defense argues that Arden Fowler did not make these statements, and therefore did not commit bribery. The defense states:

This text message screenshot is fake.

Given how common these types of attacks are, politicians like Arden Fowler are required to use a new, special messaging app that helps neutralize disinformation campaigns. The messaging app Governor Fowler uses allows chat participants to edit anything in the chat: the messages the Governor sent, the messages other people sent to the Governor, and all metadata associated with a message, like the time of day a message was sent. Anyone using this messaging app can create content or edit content and produce a transcript that is indistinguishable from the original transcript.

The following text message screenshot was then given to jury members:

[Screenshot (see Figure 3)]

Although this message looks like it was sent by Arden Fowler, it wasn’t. It was fabricated using the transcript editing ability of the text messaging app Governor Fowler uses, as an example of how fake messages can be created.

[R1–R3*]

□ *(R1-d) The screenshot is fake because anyone can create real-looking fake messages using this app, just like the one above about space walking

---

5—TOOL-TRYABLE

The defense argues that Arden Fowler did not make these statements, and therefore did not commit bribery. The defense states:

This text message screenshot is fake.

Given how common these types of attacks are, politicians like Arden Fowler are required to use a new, special messaging app that helps neutralize disinformation campaigns. The messaging app Governor Fowler uses allows chat participants to edit anything in the chat: the messages the Governor sent, the messages other people sent to the Governor, and all metadata associated with a message, like the time of day a message was sent.

Anyone using this messaging app can create content or edit content and produce a transcript that is indistinguishable from the original transcript.

Governor Fowler’s lawyers then offered jury members a chance to use a version of the text messaging app that Governor Fowler uses. Feel free to use the text message editing tool below to see how easy it is for someone to make anyone say anything.

[Interactive GUI (see Figure 4)]

[R1–R3*]

□ *(R1-d) The screenshot is fake because I can create fake messages that look real too

---

6—TOOL-PROACTIVE

The defense argues that Arden Fowler did not make these statements, and therefore did not commit bribery. The defense states:

This text message screenshot is fake.

Given how common these types of attacks are, politicians like Arden Fowler are required to use a new, special messaging app that helps neutralize disinformation campaigns. The messaging app Governor Fowler uses allows chat participants to edit anything in the chat: the messages the Governor sent, the messages other people sent to the Governor, and all metadata associated with a message, like the time of day a message was sent. Anyone using this messaging app can create content or edit content and produce a transcript that is indistinguishable from the original transcript.

Governor Fowler and Fowler's team made preparations, in advance, to protect against these types of hacks. To prove that anyone can create text messages "supposedly" sent or received by Governor Fowler, the Governor's IT department is constantly sending fake text messages among themselves. For example, the below message from September 16 notes Governor Fowler's intent to travel to the Moon.

[`Screenshot (see Figure 3)`]

[`R1-R3*`]

□ *(R1-d) The screenshot is fake because the original message about signing bills for money could have been one of the fake messages sent by the Governor's team

---

You have now seen the most important pieces of evidence in this case. The prosecution and the defense do not feel you need any more evidence to make your final decision—whether the Governor accepted a bribe.

Arden Fowler is:

□ Guilty of accepting a bribe
□ Not guilty of accepting a bribe

Which evidence or other factors did you consider when deciding whether Governor Fowler is guilty? [free-text]

In this study, you previously said:

[`<R2> belief answer`]

(If 'Yes' or 'Maybe') **Is there any evidence that would have led you to change your mind about whether Arden Fowler took a bribe?**

□ Yes
□ No
□ Maybe

What evidence might help change your mind? [free-text]

Please answer the following questions on a scale from 1 (strongly disagree) to 7 (strongly agree).

1) I'm pretty good at working out when someone is trying to fool me.
2) I'm usually quick to notice when someone is trying to cheat me.
3) I'm pretty poor at working out if someone is tricking me.
4) I quickly realize when someone is pulling my leg.
5) It usually takes me a while to "catch on" when someone is deceiving me.
6) I'm not that good at reading the signs that someone is trying to manipulate me.
7) My family thinks I am an easy target for scammers.
8) If anyone is likely to fall for a scam, it's me.
9) My friends think I'm easily fooled.
10) Overall, I'm pretty easily manipulated.
11) People think I'm a little naïve.
12) I guess I am more gullible than the average person.

Please answer the following demographic questions.

What is your age? [free-text]

Please specify the gender with which you most closely identify.

□ Male
□ Female
□ Other [free-text]
□ Prefer not to say

Please specify your ethnicity.

□ White
□ Hispanic or Latinx
□ Black or African American
□ American Indian or Alaska Native
□ Asian
□ Other [free-text]
□ Prefer not to say

Please specify the highest degree or level of school you have completed.

□ Some high-school education
□ High-school education or equivalent
□ Vocational training (e.g., NVQ, HNC, NHD)
□ Associate's degree (e.g. AS, AB)
□ Some college/undergraduate education; no degree
□ College/undergraduate degree (e.g., BSc, BA)
□ Graduate/postgraduate degree (e.g., MSc, MA, MBA, PhD, JD, MD) – please specify [free-text]
□ Other [free-text]

In general, how would you describe your political views.

□ Very conservative
□ Conservative
□ Moderate
□ Liberal
□ Very liberal

How frequently do you give computer or technology advice (e.g., to friends, family, or colleagues)?

□ Often
□ Sometimes
□ Rarely
□ Never

Have you ever faked a screenshot, maybe as part of a joke?

□ Yes
□ No, but I could if I wanted to
□ No, and I wouldn't know how to even if I wanted to
□ No, it is not possible to fake a screenshot

Appendix B. Qualitative Coding Codebook

Qualitative codebook used to code **Survey 1** responses to the question: How did the defense's evidence impact your belief?

Code	Comment
Opposite of intended effect	Participant, because of the deniability evidence, is now leaning (more) toward disbelieving the defense. Must be a very clear statement.
No effect	Participant clearly indicates they had no reliance on the evidence, or expresses complete disagreement with the defense.
Ambiguous (either no effect or small effect)	Answers that could reasonably be considered "no effect" or "small effect." Statements using cautionary terms—it did not very much, not a big influence, or did not have much of an impact—were coded into this category.
Small effect	Participant clearly indicated an effect, but it was small or contingent. Statements using terms like—just slightly, just a bit, and very little—were coded in this category. If the participant simply repeated the defense's evidence, we assumed the evidence had a "small" impact.
Large effect	Participant was greatly affected. Participant may express some amount of doubt; modifier used implies significance (e.g., participant uses bold language such as being "almost positive" of something)
Changed belief (from guilty to either equally likely or not guilty)	Participant clearly indicated a change of heart. Must be from either guilty to not guilty or guilty to on the fence.
Other	Participant's statement cannot be reasonably interpreted to be any of the other codes.