On the Credibility of Deniable Communication in Court

JACOB LEIKEN and SUNOO PARK, New York University, USA

Over time, cryptographically deniable systems have come to be associated in computer-science literature with the idea of "denying" evidence in court — specifically, with the ability to convincingly forge evidence in courtroom scenarios, and relatedly, an inability to authenticate evidence in such contexts. Indeed, in some cryptographic models, the ability to falsify mathematically implies the inability to authenticate. Evidentiary processes in courts, however, have been developed over centuries to account for the reality that *evidence has always been forgeable*, and relies on factors outside of cryptographic models to seek the truth "as well as possible" while acknowledging that all evidence is imperfect. We argue that deniability does not and need not change this paradigm.

Our analysis highlights a gap between technical deniability notions and their application to the real world. There will essentially always be factors outside a cryptographic model that influence perceptions of a message's authenticity, in realistic situations. We propose the broader concept of *credibility* to capture these factors. The credibility of a system is determined by (1) a threshold of quality that a forgery must pass to be "believable" as an original communication, which varies based on sociotechnical context and threat model, (2) the ease of creating a forgery that passes this threshold, which is also context- and threat-model-dependent, and (3) default system retention policy and retention settings. All three aspects are important for designing secure communication systems for real-world threat models, and some aspects of (2) and (3) may be incorporated directly into technical system design. We hope that our model of credibility will facilitate system design and deployment that addresses threats that are not and cannot be captured by purely technical definitions and existing cryptographic models, and support more nuanced discourse on the strengths and limitations of cryptographic guarantees within specific legal and sociotechnical contexts.

1 INTRODUCTION

Deniable encryption, and cryptographically deniable systems more broadly, refer to privacy-protective communication systems in which parties can convincingly "deny" their past communications to a powerful coercive adversary. Originally proposed in the 1990s as a theoretical concept [6], deniable communication has become highly practical since, and reached billions of users through popular messaging applications including WhatsApp and Signal [29, 12].

Over time, cryptographic deniability has become increasingly associated with the idea of "denying" evidence in court — more specifically, of (1) the *ability to forge evidence* in courtroom scenarios, and relatedly, (2) an *inability to authenticate evidence* in such contexts. Some prior literature on cryptographic deniability has suggested that, if deniability is implemented properly, courts should be unable to utilize any evidence derived from deniable communication systems — and even further, that if courts treat evidence from such systems as credible, then the purpose of deniability is undermined [12, 45]. Such literature also suggests that if judges fully understood deniability, they might not allow evidence from deniable communications to be admitted [12, 45]. Of course, not all literature on deniability ties the success of deniable communication this closely to courtroom applications; however, the association is prominent, with about half of relevant recent papers mentioning courts and/or prosecution as a threat model that deniability addresses.[1]

Mathematical definitions of deniability do directly imply an ability to forge[2] evidence (i.e., (1) above). Within many of the models considered in the deniability literature, if deniability is properly implemented, then there is a mathematical guarantee that any message someone claims to have sent or received is just as likely as any other message to have

[1] Seven of the sixteen papers mentioning cryptographic deniability on the IACR's Cryptology ePrint Archive [22] from January 1, 2020 to June 1, 2025 mentioned courts and/or prosecution as a threat model that deniability addresses. The original paper proposing deniability [6] did not mention courts or prosecution.

[2] In this paper we do not use the word "forgery" and its variants to refer to a specific common law crime (as codified, for example, in 18 U.S.C. ch. 25). Rather, we use the term with its colloquial definition: "The act of forging something, especially the unlawful act of counterfeiting a document or object for the purposes of fraud or deception" [36].

Authors' address: Jacob Leiken, jrl9854@nyu.edu; Sunoo Park, sunoo.park@nyu.edu, New York University, New York, New York, USA.

been the one actually sent or received, even given the cryptographic transcript. Applying this reasoning, treating any particular message as more credible than another would be unfounded (i.e., (2) above). Indeed, within these mathematical models, (1) the ability to forge messaging history may *logically imply* (2) the inability to authenticate messaging history.

But in almost any realistic context, the ability to forge messaging history does not imply an inability to authenticate messages, when other contextual and circumstantial evidence and possible witnesses are taken into account. That is, the above logical implication holds only within a mathematical model that strictly limits the information available.

In our view, the deniability literature's focus on courtroom applications is misplaced. Even if deniability is implemented perfectly, we expect evidence derived from deniable communication systems to be often relied upon in courts — not because of a lack of understanding of deniability, but as part of the proper operation of evidentiary systems.

Courts authenticate evidence using processes that have been developed over centuries to account for the reality that *evidence has always been forgeable*. These processes are heavily reliant on circumstantial evidence and witnesses, and deliberately admit evidence that could conceivably have been forged — such as verbal testimony, letters, photographs, and so on — for evaluation by a jury or judge (who may eventually decide they were actually faked). Just like chat transcripts, these other types of evidence are generally admitted in legal proceedings absent strong evidence of their unreliability, which must go well beyond the mere possibility of forgery.

None of this means that cryptographic deniability is useless; in our view, courtrooms were never the killer application for cryptographic deniability. Moreover, deniability may play an important role in limited contexts within legal proceedings, even if the fact of deniability by itself should not be grounds for evidence to be excluded from a courtroom. In routine use, deniable communication preserves a stronger ability to argue against evidence that lacks strong authentication (i.e., as noted above, most evidence), compared to other more attributable communication methods. As such, the use of deniable communication places a renewed emphasis on traditional fact-finding processes within the legal system, which were designed to deal with unreliable and forgeable evidence, shifting the focus to questions such as: is the witness who brought the transcript to the court reliable? How was the evidence discovered? Is it corroborated by other parties, and what are their incentives? Is the transcript consistent with the relevant party's other conduct?

Deniability could be an explicit issue in a legal proceeding if a party raises a challenge to evidence that they can back up with a concrete reason to believe that that particular evidence has been modified or forged — again, beyond the mere possibility of forgery. This need not be based on technological expertise; it could be as simple as a party to the conversation claiming that they never received the message in question, and perhaps backing this up by pointing to the message history in their app. Such a claim would then lead to a more detailed investigation of the evidence, in which technological expertise on deniability might well be relevant. If so, there are existing, if imperfect, processes in the legal system to bring in the testimony of experts.

Just because evidence could be forged does not mean it has been. Indeed, if no evidence of a type that could in principle be forged were admissible in court, then pretty much no evidence could be admitted. Verbal testimony would be prohibited because people can lie, essentially all documents would be excluded because they can be faked, and items found at the scene of a crime would be inadmissible because they could have been planted. Even cryptographically signed communications can be forged if someone steals the signing key.

The intended role of evidence law is to ensure accurate fact-finding by, among other mechanisms, filtering out evidence that is clearly irrelevant or prejudicial in certain ways[3] — and to admit the rest, even though it may be unreliable, for the jury or judge to form their own opinion and their own doubts. Although existing evidentiary

[3]For example, a victim's past sexual behavior is inadmissible in cases against their accused abusers. Fed. R. Evid. 412. There are other reasons prejudicial evidence and limited other types of evidence could be excluded, which are beyond the scope of this paper.

frameworks are far from perfect, they have been honed over hundreds of years in an effort to administer justice in an environment where almost all evidence is unreliable, the risk of forgery is omnipresent, and these inconvenient facts often present real practical problems (as in the example below). In this sense, we see deniability as furnishing an additional example of a old — and challenging — problem with which the legal system is well acquainted, rather than introducing problems that newly disrupt or confound the existing system.

"Deniability" in court in the Cold War. The high profile Cold War case of Alger Hiss, from the 1940s, is one example where the possible forgery of evidence featured prominently. The case elicited extensive commentary on the Anglo-American evidentiary system and the merits and drawbacks of admitting evidence known to be possibly unreliable [37]. Hiss strenuously argued over the course of two trials and an appeal that the government's evidence against him was insufficient. He claimed that they key evidence in the case — contested letters — were forgeries written on different typewriters. Despite contrasting testimony from experts, the letters were introduced as evidence, Hiss's conviction was upheld, and over protracted discussion in the following years, the legal profession decided against making any changes to the system that allowed the contested letters to come into trial [37]. To this day, it is unclear whether the letters were forged. In the absence of certainty, it was determined that admitting the letters was the right thing to do. It allowed the jury to weigh the testimony of Hiss, other witnesses, and experts, to select the verdict they believed reflected the truth. That two juries entered the same verdict, despite different testimony and different presentations by the defense and prosecution, supports the claim that this system works as designed in the face of uncertainty.

Credibility, not deniability. Acknowledging that there are almost always factors outside of cryptographic models that influence perceptions of an artifact in practice,[4] we introduce a new model of *credibility*, comprised of three elements. The first is the *threshold of believability* that must be surpassed for a forgery to be taken as authentic by a relevant viewer in a specific sociotechnical context and threat model; while this is generally not precisely quantifiable, it is nonetheless useful to identify explicitly and reason about. Given this threshold, then, credibility is determined by two related axes: the *ease of forgery* within the system and the default *retention* policy of the system or conversation.

We see credibility as an essential and overlooked component of threat modeling for deniable communications, in realistic contexts not limited to courtroom scenarios. As such, we hope that our conceptual framework can guide the design and implementation of secure communication systems in practice, to facilitate the design of technologies better to equipped deliver on the promise of deniability in diverse sociotechnical contexts.

Outline and contributions. Section 2 discusses related work, and Section 3 provides background on deniable communication. Sections 4 and 5 present our novel contributions. In Section 4, we introduce the new conceptual framework of *credibility* and explain its implications for deniable communication within and beyond the courtroom. In Section 5, we analyze how evidence derived from deniable communication systems is likely to be treated in court, including an overview of existing mechanisms in US legal systems that are meant to deal with unreliable or forged evidence (including complex evidence of a technical nature), and a discussion of these mechanisms' likely application to cases involving deniable communication.[5] Section 6 offers further discussion and Section 7 concludes.

[4] See, e.g., https://www.xkcd.com/538.
[5] While our analysis in this section focuses on the US system, many aspects are generalizable. In particular, the general principles underlying rules of evidence and the idea that unreliable or contested evidence may be excluded are reflected in many court systems.

2 RELATED WORK

In the last two years, four studies conducted by computer scientists aimed to determine legal and/or social perceptions of deniability [45, 12, 41, 40]. The next two subsections discuss these prior works and how they relate to this paper.

Deniability in Courts. Yadav et al. [45] and Collins et al. [12], searched judge-written orders and decisions in court cases in the United States and Switzerland respectively for situations where evidence relating to secure messaging was referenced. To the expressed surprise of both sets of authors, the judges writing the surveyed orders unquestioningly allowed the introduction of evidence from messaging platforms which have cryptographic deniability features. We searched legal databases using the queries developed in Yadav et al. [45] for the time period since the piece's publication and found that the trend has continued. The tone both studies use when discussing this finding is notably crestfallen, with Collins et al. stating:

> "Our study, along with with Yadav et al.'s findings show that deniability seems irrelevant in court cases in Switzerland and the United States. Although the results cannot be generalised to other countries, they provide evidence that cryptographic deniability does not translate into real-world deniability and that a purely cryptographic approach does not impact the legal setting. When one party claims a forgery, judges consistently reject these claims." [12]

Yadav et al. takes an even more negative tone; its discussion section contains paragraphs labeled: "Deniability is ineffective," "Deniability is hard to achieve in legal cases," and "Even a screenshot is not deniable." [45] These two sets of researchers, with the knowledge and expertise of cryptographers, rightly understand that a message displayed on a platform which has implemented deniability can never be cryptographically verified as legitimate; nor, typically, can a screenshot. They may moreover be right that not every judge will understand this. From this perspective, it can be surprising or disappointing that judges are willing to permit chat transcripts or screenshots to be presented as evidence, and such an approach may appear correlated to a lack of technical understanding.[6] However, we argue that this approach reflects the correct functioning of the evidentiary system, even — or perhaps all the more so — given a full understanding of the cryptographic guarantees of deniability.

Furthermore, Yadav et al. write that "[d]eniability *requires* social and legal acceptance to be effective" where "legal acceptance means that the courts accept deniability and may not consider messages as evidence." [45] These statements make explicit a sentiment present in many earlier quotes: that deniability is "ineffective" unless judges exclude evidence from deniable communication systems from court. In contrast, this paper argues that deniability is valuable even if it is not treated as grounds to throw evidence out of court; that indeed, the value of deniability lies largely elsewhere.

Perceptions of Deniability. Reitinger et al. [41] and Rajendran et al. [40] developed prototypes of a Signal-like messaging service with cryptographic deniability built-in. They surveyed potential users for understanding and opinions on the functionality and desirability of such a feature. Reitinger et al.'s experimental setup involved participants who were asked to act as if they were potential jurors in a politician's bribery trial, while Rajendran et al. surveyed participants on whether they would want or utilize this feature in their own messaging.

These works sought to determine how users perceived deniability, how to best educate people on how and why to use deniability, and to estimate people's preferences, respectively in legal and social contexts. Our work has the distinct and complementary motivation of examining how deniable communication is and should be treated under the law of

[6] "We need to raise awareness with legal stakeholders that deniability allows bad actors to forge messages and use them in court since courts accept WhatsApp chat [*sic*] as evidence." [45]

evidence. The question of how potential jurors are likely to understand electronic communication evidence (with or without deniability), in particular, is closely related to our topics of interest. A better understanding of this question could influence how cases involving this type of evidence are litigated, and how such evidence is presented in courts.

Both studies found that before participants were given explanations about deniability, they were very likely to believe chat transcripts presented to them were authentic [41, 40]. However, both found non-cryptographic, non-legalese techniques to get users to question chat transcripts. The simplest technique was to allow users to interact with a user interface that had implemented deniability in a way intended to enhance conceptual understanding [41, 40].

Other Related Works. A breadth of research across varied disciplines has demonstrated that individuals communicate and understand communication differently depending on the medium and its perceived attributes. Academic marketing literature addressing this topic, for example [34], notes that the effectiveness of a message can vary widely by medium because of changes in the perception of the recipient. The communication technology in use changes the way individuals communicate to make decisions, making a group more or less collaborative [26]. Intuitively, users are less likely to share information that can be categorized as "intimate" if they are communicating a channel that they perceive to be open or insecure [16]. Individuals tend to change their behavior when they believe they are being surveilled, whether or not they are in fact [5], and to speak less freely when they know their communication can be monitored [15].

Many privacy theorists struggle with a "privacy paradox," that individuals often choose to share more information about themselves publicly as communications technology has advanced, a decision in apparent conflict with their stated value for privacy [1, 21]. Some prominent theories of privacy suggest that at least some of these apparent conflicts may arise from conflating different notions of privacy [42, 33]. Other literature has shown that users tend to be unaware of data practices and don't often know what data is being collected about them and for what use [43], and that it is onerous or impractical to maintain such detailed knowledge [27, 30]. Voluntary disclosure aside, individuals still view a choice to disclose information differently from (risk of) exposure that is inadvertent or against their will [21].

The above literature shows that individuals communicate and understand communications differently based on perceived security and privacy properties of the communication medium. A diversity of options allows individuals to choose the technology most suited to purpose, and deniability occupies an important space in that spectrum.

3 DENIABILITY

What is deniability? The simplest meaning of "deniable" is "possible to deny," and the verb "deny" is defined commonly as to declare untrue or refuse to believe or acknowledge (e.g., [35]). The dictionary definition we found most relevant to the context of deniable communication within computer science is "being such that plausible disavowal or disclaimer is possible" [35]. Technical definitions within computer science consider the disavowal or disclaimer of either the content of a communication (*transcript deniability*),[7] or its authorship or provenance (*conversation deniability*),[8] often by providing plausible evidence that a different communication took place instead. Broadly, the goal of deniable communication systems is to eliminate or reduce the availability of cryptographic evidence that would provide convincing documentation of the contents or provenance of an electronic communication. As further discussed below, the precise technical definitions proposed to achieve this goal vary, as do the adversaries and threat models they address.

A walk in the woods. A commonly invoked scenario to illustrate the ideal of deniable communication is a "walk in the woods." A conversation held during a walk in the woods is ephemeral, as is the mere fact that the conversation

[7] As in *deniable encryption*, e.g., [6].
[8] As in *deniable authentication*, e.g., [10].

occurred. As such, the contents of the conversation can only be repeated through the individuals' memory. Whether or not a third party believes that the conversation occurred, or the relayed contents, depends in practice on their view of the trustworthiness of the person repeating the conversation.[9]

In contrast, electronic communication often creates long-lived evidence trails distributed across multiple parties including non-parties to the communication (such as intermediary platforms), which can systematically support convincing and detailed documentation of past communications.[10] This can be the case even for communication systems that support strong privacy guarantees of other types, such as end-to-end encrypted (E2EE) systems. This is not an oversight, but rather, simply not the type of threat that E2EE is designed to protect against.

A key motivation of deniable communication is to create electronic communication systems that have similar deniability properties to the walk in the woods. Why? Secure and private communication, including deniable communication, is valuable for privacy and freedom of expression.[11] Deniable communication techniques can also be valuable as a building block for other security- or privacy-enhancing systems.[12] Furthermore, research on deniable communication can enhance our understanding of the theoretical and practical boundaries of cryptographic techniques. It is debatable whether an electronic communication system could ever achieve deniability commensurate to in-person communication, and it is an interesting open question how close we can get both in theory and in practice.

Deniable communication primer. Standard secure communication often involves two components: (1) *authentication*, where senders attest that they are the originators of the messages, in a way that recipients can verify, and (2) *encryption*, where senders encrypt content before sending, in such a way that only intended recipients can decrypt it. Both these properties rely on the secrecy of some information known only the sender or recipient, respectively, called a *secret key* (much like a password). By using their secret key, a sender can attest that they wrote a message by *cryptographically signing* it, and they can decrypt messages addressed to them. Nobody else, without the relevant secret key, has these capabilities. But if an adversary compromises a secret key, then all bets may be off: they may be able to impersonate and decrypt messages intended for the victim. A cryptographic transcript and the associated secret keys would typically constitute pretty solid evidence both of who participated in a conversation and what they said.[13]

Deniable encryption, then, adds a layer of security on top of standard encryption, to provide protection even against a strong coercive adversary who can demand access cryptographic transcripts and secret keys from conversation participants, by allowing the participants to generate alternative secret keys that look consistent with the cryptographic transcript of the conversation, but make it look like either (1) something else was said (*transcript deniability*, which relates to encryption) or (2) they did not participate in the conversation (*conversation deniability*, which relates to authentication). For example, if Alice and Bob are using an encryption scheme which has implemented cryptographic deniability, the result of Eve's "decryption" could be any string of Bob's choosing, or a decoy string set in advance.

[9] A party could surreptitiously record the conversation, and this has gotten progressively easier with modern recording technology. But importantly, the "default" way that a walk-in-the-woods conversation happens is without surreptitious recording. That is, a party must go out of their way in order to make and conceal a recording, and moreover must do so in the moment as it becomes impossible after the fact. In this paper, when we discuss the walk-in-the-woods ideal, we generally assume there is no surreptitious recording.

[10] Such evidence is not irrefutable, as we discuss more later. For example, an adversary who compromised a user's password could send messages in the user's name that would pass cryptographic verification, but would in fact not have been sent by the user. Our point here is that such evidence trails are quite reliable in general, and available systemically.

[11] For example, research has shown that people speak less freely when they know their communication can be monitored [15], and the UN Office of the High Commissioner on Human Rights has emphasized that strong encryption is essential to fundamental human liberties including free expression [23].

[12] The original proposal of deniable encryption suggested two fairly specific cryptographic applications: receipt-free voting protocols and incoercible multi-party computation [6]

[13] Someone who had both of these could use the keys to decrypt the transcript themselves — and typically, without deniability features designed specially for the purpose, it is not feasible to decrypt encrypted transcripts in multiple plausible-looking ways.

Cryptographically deniable systems are already in widespread use. Most notably, the Signal protocol, used by the popular end-to-end encrypted WhatsApp and Signal messaging apps as well as other secure messaging systems, implements conversation deniability [44]. This allows conversation participants to deny their participation (conversation deniability), but not to deny the contents of specific messages (transcript deniability).

If two people are communicating using an encryption scheme that implements (either kind of) cryptographic deniability, this precludes any cryptographic evidence that a specific message has been sent by a specific person [6], *either* because the conversation *or* because the transcript is deniable. If Bob ever leaks a message from Alice, Alice has a mathematical way of repudiating that message. The mere possibility that any participant can repudiate a message in this way shows that Bob *could* be lying. Thus, conversations over deniable encryption bring us a step closer to the "walk in the woods." With message deniability, for example, there may be evidence that a conversation occurred (Eve could have watched Alice and Bob walk into the trees), but proving what was said has become much harder. Eve may eventually think that she knows what was discussed and who said what, but Eve's impression would have to be based on interviews with Alice and Bob and her perception of their trustworthiness.

Variants of deniability. A range of variant definitions of deniability exist. For this work, we construe the term broadly. As our work is largely agnostic to definitional variants, we refer to [7, §1.5] for an overview of variation in cryptographic definitions. We highlight one non-cryptographic deniability notion, as it relates to our later discussion of credibility: *hidden volumes* constitute a form of deniability in the context of operating systems. The now-defunct application TrueCrypt provided a fairly mainstream implementation of this [2].[14] It allowed users to type a preset fake password into a locked computer, revealing a "safe" partition and hiding the original [11]. This allowed users to deny the existence of files saved on their device. In this case, a user acknowledged that a device was their own, but effectively denied the existence of whatever was on the hidden partition.[15]

4 CREDIBILITY: BRIDGING CRYPTOGRAPHIC DENIABILITY AND REAL-WORLD ATTRIBUTION

Cryptographic deniability is an important property with precise mathematical definitions. But as discussed above, for reasons that lie beyond the scope of cryptographic models, even perfectly implemented strong deniability does not render communications impossible to attribute in practice.

We propose a new framework to address this gap. Our new definition of *credibility* breaks down the key sociotechnical and contextual considerations that influence the imperfect art of attribution in real-world contexts. Cryptographic deniability and other technical system properties are important factors that influence credibility; at the same, credibility is an inherently sociotechnical concept, and cannot be analyzed based on technical properties alone.

Analyzing communication systems through the lens of credibility make explicit the sociotechnical considerations that impact how "denials" are enacted and received in practice, providing a much-needed language to reason about critical aspects and likely outcomes of deniable system use, which are typically considered "outside the model."

Next, Section 4.1 elaborates on the gaps we seek to bridge, Section 4.2 presents our credibility framework and its three elements, and Section 4.3 provides broader discussion on our framework.

[14] The application is said to have been downloaded 30 million times over the 10 years it was active [39].
[15] This is not a true implementation of deniability. Rather, it is referred to as "false bottom" encryption. As far as we are aware, true deniability has never been implemented for data at rest. Deniability for data at rest is generally outside the scope of this paper.

4.1 Three gaps to bridge

Gap #1: Is the record preserved? Written (or otherwise recorded) communication is inherently less repudiable than an oral conversation due to the creation of the record. Consider the example of a handwritten letter. Anyone (with a writing sample, time, and patience) could produce the exact same letter, yet in many realistic contexts, the document would be considered stronger evidence than an oral retelling.

Deniable communications schemes are likewise less repudiable than a walk in the woods. If Bob shows Eve a chat transcript on his phone of a conversation between Bob and Alice, there may be no proof that Alice actually sent the displayed messages, but the display in front of Eve may well give her more confidence in their veracity and detailed accuracy than Bob orally relaying a conversation. This intuition reflects the work required by Bob in both states. If the transcript were real, then there would be a record that looks exactly like this which would be easy for Bob to show Eve. If it were fake, Bob would have had to undertake some effort to create the fake, thereby diverging from the default condition of the system. This state of affairs contrasts with a walk in the woods, where the work required to relay the true or forged contents of the conversation is effectively identical.

Similarly, if Alice and Bob communicated over a platform implementing cryptographic deniability and Eve is a cryptographer who is in a position to examine Bob's device memory, then if the secret keys and randomness are on Bob's device at the time of Eve's (surprise) examination, Eve may be more likely to believe them as the originals. Although Bob could in principle have prepared fake secret keys and randomness and left them on his phone for Eve to find, this is not the default state of affairs; Bob would have had to go out of his way to prepare the fakes. As such, Eve may or may not believe it is more likely that the secret keys and randomness are unmodified, depending on the context.

In all these cases, the starting point is the original records proffered, records which remain intact unless someone deletes them or goes out of their way to do tampering or forgery: the letter, the message, or the cryptographic material. The field of *forensic science* encompasses scientific approaches to investigating and analyzing such records and their traces which may be used as evidence in legal proceedings (e.g., [32, 3]). Broadly, forensic science concerns all kinds of evidence, from handwritten letters to blood splatters and much more. Digital forensics is the sub-field that focuses on digital evidence and investigations (e.g., [9, 25]). Successfully passing off a forged letter or a forged Signal transcript as authentic is much more difficult under forensic examination, than towards casual observers; and where a plausible doubt is raised about a particular piece of evidence in court, forensics experts are likely to be called in.[16]

The walk in the woods mitigates this issue by minimizing the possibility of any original records of the communication being available at all. Electronic communication media generally leave more traces than a literal walk in the woods; nonetheless, in a similar spirit, we can design systems that minimize retention of records. A common example available on many popular messaging platforms is "disappearing messages."[17] A more thorough alternative might involve constantly wiping keys and refreshing randomness in device memory. Indeed, the Signal protocol does something similar [29];[18] that said, simple deletion in software is unlikely to withstand forensic examination in practice, and secure deletion is very difficult to get right [24, 14]. These types of features are not cryptographic in nature,[19] so naturally lie outside cryptographic deniability definitions; they relate to system design and defaults, and are still within the scope of security engineering.

[16]While this is not a very high bar, we note again that the threshold is to cast concrete doubt on the authenticity of a specific piece of evidence (e.g., this particular letter), and pointing out the possibility of forgery of evidence (e.g., letters in general) in a given medium, without more, is generally not enough.
[17]"Disappearing messages" aren't perfect in practice, but they shift defaults of retention in an impactful way [19].
[18]Signal deletes per-message keys upon decryption.
[19]It is not currently possible to prove non-retention cryptographically. Recent papers have suggested quantum techniques to do so, but these are not practically feasible at the moment [4].

Gap #2: How much effort? The preceding hypotheticals note not only that Bob would have to go out of his way to forge records, but also the effort that Bob would have had to invest in so doing. Existing definitions tend to focus on whether or not forgery is possible, and treat forgery as "easy" if there exists an efficient algorithm to perform it. In practice, the perceived likelihood of forgery may depend on how much effort is required. While the concrete efficiency of the algorithm is one relevant consideration, many other practical considerations are non-technical or only partly technical in nature. Do forgery tools come with the communication platform, or would they have to be obtained separately? Are there existing libraries or applications that perform forgery, or would one have to write the code oneself? Does forgery require technical expertise? Are existing tools easy to use? For a coder, or for a layperson? Could a layperson pay an expert to do the forgery instead? How much would that cost, and how hard would it be for them to find a competent and discreet expert?

Gap #3: Who is trying to convince whom, in what context? Finally, context and relationships matter in whether someone is likely to be believed. For example, if Eve knows Bob is a serial liar who has faked letters before, she will be less likely to believe him (even if provided a detailed communication transcript). If Bob is a long trusted friend or family member of Eve, then she will be more likely to believe him (even without a detailed communication transcript). In either case, if Alice is also friends with Eve and corroborates Bob's story to Eve directly, this may boost Bob's credibility; if Alice says Bob is lying, this may damage his credibility. Incentives and content are critical contextual factors. If Alice is Bob's creditor and she corroborates that she released him from a debt, that is especially credible. A chat transcript between politicians about matters of national concern, or a chat transcript that appears to be evidence of a crime, will receive much more scrutiny than personal messages of no public concern that do not make anyone look bad.

The three gaps that we identify highlight a range of contextual "outside-the-model" factors that feature in realistic assessments of credibility — whether or not cryptographic attribution methods are available. The importance of contextual factors in assessing credibility is naturally heightened in those situations where cryptographic attribution is not available — that is, most situations, as most kinds of evidence have no cryptographic attribution.

Within a cryptographic model, if messages can in principle be forged, there may be nothing more that can be done to ascertain the likelihood that a given message has been forged — that is, forgeability may imply impossibility of attribution. But outside of the cryptographic model, as underscored by the discussion above, it is essentially never the case that there is nothing more that can be done to ascertain the likelihood that a specific message in a specific context has been forged. Neither Eve, nor the judge in the courtroom scenario, needs to pack up and go home upon discovering the possibility of forgery; instead, they turn outside the cryptographic model, to contextual factors that they would have considered in any case, but are now their primary recourse. And indeed, the contextual factors highlighted above are the sort of things that could well feature in courtroom scenarios trying to assess the reliability of evidence.

4.2 The three elements of credibility

Crystallizing the above discussion, communications technologies can be seen as "implementing" credibility on two axes.

The first is message **retention** (related to Gap #1). The less time a message is retained by default, the more often it will be credible it will be for message participants to say that they no longer have the original. As discussed in Section 4.1, retention is outside the scope of cryptographic deniability.

The second is **ease of forgery** (related to Gap #2). If it is easy to create a false message that will be believable as a possible original message, then the mechanism implementing deniability also benefits credibility. Cryptographic

deniability focuses entirely on this axis. As noted in Section 4.1, technical features such as cryptographic deniability are necessary but not sufficient to determine ease of forgery in practice, as other contextual factors also matter.

The third factor of our credibility framework is the **threshold of believability** (related to Gap #3): that is, the bar which a forgery must pass to be likely to be believed in a relevant sociotechnical context. As we explain further below, unlike the preceding two axes, the threshold of believability is independent of the communication medium.

The following subsections provide more detailed discussions on the three elements and how they fit together.

4.2.1 Retention. The retention of past communications refers to the default amount of time a message is retained after viewing. This default could be either at the service level or the conversation level. Any communication could be recorded without consent, such as a surreptitious audio recording of a conversation during a walk in the woods or a screenshot of a message sent in Snapchat. However, these require additional effort in the moment, meaning that parties to the communication would have to be actively adversarial at the time of the communication. This axis instead measures the default where the parties communicating are non-adversarial. Retention varies greatly, with one end being no retention past the initial viewing (i.e. a walk in the woods, Snapchat), and the other being permanently retained by many parties (i.e. an official government edict).

Disappearing messages are sometimes not included in discussions of deniability because they are not cryptographic. But retention is an important aspect of secure system design, in part because disappearing messages can mirror many of the characteristics of the prototypical walk in the woods. Assuming messages are "disappeared" thoroughly, which comes with its own host of technical challenges, a disappearing message also means that there is, by default, no record of a conversation after it happens. The Signal protocol implements deniability through a short-retention scheme [29], so significant thought has gone into this axis in practice, albeit not in this explicit framing.

4.2.2 Ease of forgery. A forgery is relatively "easy" if it can be easily done by an untrained individual using readily available resources or cheaply done by a trained individual — either means that it is widely accessible.[20, 21] One end of this axis is deniability by nature or forgery that is as easy as telling a lie (e.g., a walk in the woods), and the other is cryptographically non-repudiable messaging with no way to edit messages (e.g., an encrypted, cryptographically signed email).[22] In between these extremes, towards the "easy" end, we place deniability implemented through an easy-to-access, easy-to-use interface built in to a communication platform (e.g., the system prototype devised in [40]).[23] While the computational efficiency of the forgery process is a relevant factor, we find that socio-technical and non-technical considerations tend to dominate the analysis of ease of forgery in practice (e.g., in all examples in Figure 1).

4.2.3 Threshold of believability. Consider the threshold for creating a credible forged photograph of Alice's head on Carol's body. Say the photograph looks believable at first glance but when zoomed in there are pixels that are clearly artifacts of an edit in Photoshop. For a personal post to a small audience on Instagram, this might pass the threshold of

[20] However, paying a trained individual creates an additional potential vulnerability outside the cryptographic model, in that the trained individual now knows about the forgery and could be a potential witness.

[21] Often, as the technical skill becomes more specialized, the cost of performing the skill as a service increases. But this is not true of deepfakes created using generative AI, which are increasingly used to create images passing a requisite threshold of believability. The technical skill required to make a convincing deepfake is fairly high, but skilled individuals currently offer to create deepfakes for relatively cheap because online platforms, at least until recently, have incentivized this behavior [28].

[22] The relevant threshold of believability may affect where on the ease of forgery axis a service falls. Consider a technology which allows for the easy creation of non-believable forgeries but which requires exponentially higher skill and time to make believable forgeries. For a low threshold, this service would be very high on this axis, but for a high threshold, it would likely be lower. Having one plot, as in Figure 1, assumes that the rate at which the quality of the forgery improves is proportional for all technologies depicted.

[23] Notably, existing applications do not currently offer a readily accessible, user-friendly interface for denying communications, even where they do have cryptographic deniability features. The example in [40] was a research prototype only.

believability. But if that Instagram post went viral and made the national news, then the heightened scrutiny it would receive would likely bring these artifacts to light. Similarly, for a law enforcement investigation or court case, this forgery likely would not pass the threshold. In other words, a forgery that suffices for one purpose may not for another. This simple fact is made significant by the reality that there is almost always scope to consider contextual factors more thoroughly in assessing an artifact's credibility — so how thoroughly people are likely to investigate contextual factors ends up a critical consideration for credibility in practice. Court cases, and especially criminal cases, present a context which is specifically designed to try to bring into consideration as many contextual factors as feasible, even at relatively high cost, as someone's life or liberty may be at stake.

We define the threshold of believability as independent of the communication medium; instead, it focuses on questions such as who is trying to convince whom, and in what context, as discussed under Gap #3 in Section 4.1. Although we have introduced it as our third element of credibility, it will often be the first point of inquiry in analyzing the credibility properties of a real-world application context, such as when designing systems. Deciding on a(n approximate) threshold of believability as a initial step allows a more structured inquiry into the two axes with respect to the chosen threshold. The relevant threshold of believability in a given context could help system designers determine where on the other two axes of *retention* and *ease of forgery* a service should be located to provide the desired level of credibility for an application context, or for individuals deciding which communications technique to use. In this sense, the threshold of believability may shade an area on the graph constructed by the two axes as a target; it is not a third axis in itself.

4.2.4 Big picture. These three elements are distinct but overlapping ways of approaching credibility, not fully separate from each other. No direction or quadrant on the axes is "better" than another; the best configuration for an application depends on its context. Identifying methods of communication as existing in this problem space is helpful both to better understand the space of possibilities, and to promote a diversity of options.

In Figure 1, we plot a variety of communication methods on the first two axes of credibility. We have opted for a two-dimensional plot as the third element, the threshold of believability, is independent of the communication medium as noted above. None of the elements are precisely quantifiable or defined by a precise metric, due to inherent qualitative and socio-technical considerations, so there is necessarily some imprecision involved in the plot. We believe the graphical representation is nonetheless helpful to convey our own approximate and comparative sense of where different communication technologies lie within the new space that our credibility framework defines.

4.3 Discussion on credibility

How does credibility connect deniability to real-world attribution? There are many different reasons credibility can be low, including easy external verification — someone may claim that the front page of today's New York Times says X, but it is generally easy for a listener to determine the veracity of that claim, so the claim is not credible if New York Times front page in fact does not say X. For rather different reasons, a tornado siren would be difficult to fake credibly, because it relies on city-, or even region-wide equipment, and should be perceived by an entire region of people at once if it in fact happened. (Both of these examples appear in the plot in Figure 1.)

Examples such as these illustrate that credibility expands on the scope of deniability in a way that captures important aspects of whether true and false claims about past communications are likely to be believed in practice, which are outside of the scope of deniability alone. As theorized in literature, deniability is intended to eliminate the possibility of the coercer ever believing they have the true plaintext, thereby making the whole exercise of coercion futile (e.g., [6]). But the mere fact of the ability to generate false secrets realistically carries little significance if the coercing party

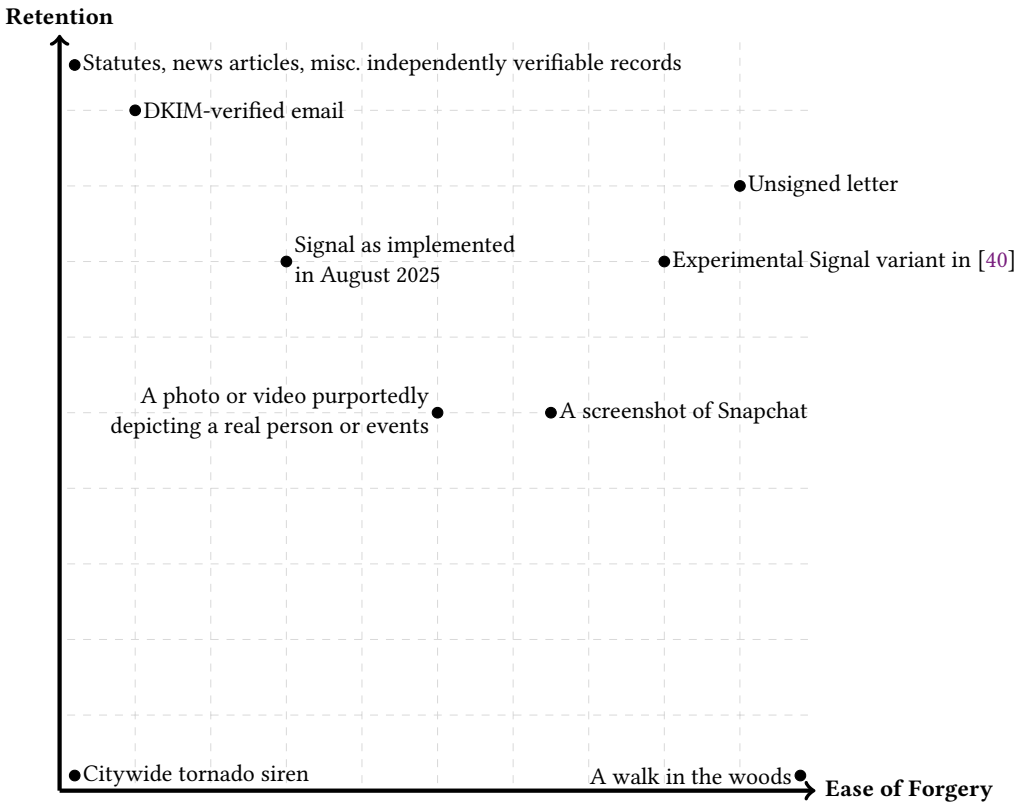

Fig. 1. Different communications techniques on the two axes of credibility

believes, for reasons outside the model, that they have a reasonable likelihood of having decrypted the true plaintext. Thus, credibility is the practical stopgap connecting purely technical notions of deniability to real-world attribution.

Put differently, deniable encryption (resp., authentication) ensures that Bob, in assessing claims made about a plaintext (resp., Alice's participation) in a disputed interaction, cannot credibly point solely to the claimed original record of communication, but rather must rely on contextual information beyond that record — and that is where credibility comes in. Deniability does not imply an inability to attribute, but rather, pushes us to look beyond the transcript by eliminating the apparent "smoking gun" of the link between the message, its encryption, and claimed keys.[24]

Why does credibility matter? As previously noted, no point in Figure 1 is the ideal; we define this concept to highlight the importance of the availability of options that mimic the full range of analogue communication.

There are many situations where a difficult to forge, permanently retained communication is the ideal. For example, the buyer of a house would be unlikely to go through with the transaction unless they were promised a difficult to forge, permanently retained contract. On the other hand, young adults developing their self expression may value or benefit from low retention, and from scenario to scenario, prefer more or less credible forgeability [13]. The related literature discussed in Section 2 suggests that developing a variety cryptographic techniques enabling digital communications

[24] As also noted above, even this link is not a true smoking gun, as keys may be compromised.

services to be provided at many points on the axes of credibility has a positive impact on the lives and welfare of everyday individuals, whether or not they are aware of the advances. Having more variety in the credibility properties offered by digital communications technologies, including the ability to mimic a wider spectrum of analog communication methods, allows individuals freer choice over proper technology for their expression.

These considerations are outside the cryptographic model, but it is still important to have a structured way of approaching them. Credibility provides such a framework for designing systems that take into account aspects of real-world threat models that would previously have been considered outside the scope of deniability.

The three elements of credibility are not comprehensive, but we believe they capture essential aspects. While no theoretical framework can include all practical considerations, we believe that the three elements we identified present a valuable approach to reasoning about credibility, which capture aspects beyond those explicitly named in the discussion above.

For example, usability is an important consideration not directly not captured by the three elements, but it interacts in subtle ways with all three. If deniability is implemented in a way that users do not understand, they may not use that feature or may use it incorrectly, and would therefore lose out on the security protections it promises. A system with poor usability may make creation of forgeries more difficult, which would be reflected under ease of forgery.

5 HOW EVIDENCE IS AUTHENTICATED IN COURT

Courts have developed a variety of mechanisms to determine what evidence can be presented at trial, within the longstanding context that most evidence introduced in courts does not have the possibility of cryptographic verification. In the US system, judges decide whether evidence can be admitted and the evidence that is admitted is presented to the jury, who may choose to take it into account in rendering a verdict. The jury does not need to view the evidence presented as inherently reflecting the truth of the matter; it is instead presented to help the jury determine what the truth of the matter is, on the understanding that the jury should do their best to determine the truth even though some of evidence presented may be contradictory or unreliable.

Consider the example of verbal testimony, which is always forgeable in principle, as people can lie. What is more, "he said, she said" situations do arise in courtroom scenarios, and in general, both witnesses are permitted to testify — even if, logically speaking, at least one of them must be lying. It is the jury's role in the system to consider the accounts presented to them and use their best judgment to reach a verdict. To try to ascertain who is right and who is wrong, and only present the one who is right, would create a recursive problem — it is the jury's role to consider all the evidence and determine who is right or wrong, and taking that away from them not only invalidates their role in the system (and the defendant's right to a jury trial, where applicable), but also leaves wide open the question of how then to reach such a determination, having thrown out the very mechanism designated by the legal system to do so. To block both stories from being presented, on the grounds that they are possibly unreliable — to exclude all verbal testimony on the grounds that it is, in principle, forgeable — could be to block what might be some of the most relevant information available to the jury, and leave them little or nothing to work with. On balance, legal systems have determined that admitting some possibly unreliable evidence, and letting the court system take it all into consideration with the knowledge that it may be unreliable, is ultimately more conducive to truth-seeking than the alternative.

At the same time, legal systems do have checks and limits on what evidence can be admitted and presented. To be introduced, evidence must satisfy an extensive list of rules defined in each jurisdiction.[25] In the US system, evidence

[25] In United States federal court, the rules governing these topics are the Federal Rules of Evidence. Each state court system has its own set of rules, but they are largely modeled after the federal rules, usually only deviating slightly [8]. Each country's court system has its own rules regarding evidence as well, but for simplicity, only the federal U.S. rules are discussed here.

must be both *authenticated*[26] and *relevant*[27] to be presented in court; the judge decides whether a piece of evidence is satisfies these requirements. When helpful, *expert testimony* may be brought in. We elaborate on these concepts below.

Finally, it is important that the legal system provides parties recourse against actual forgeries that others may attempt to use against them in court.[28] We do not believe the permissiveness of the evidentiary system generally means that parties are likely to be without recourse against actual forgeries — at least not any more for deniable communications than for other types of forged evidence such as false verbal testimony, handwritten letters, doctored images, or cryptographically signed messages forged by password hacking. If one party forges a deniable communication, in practice, the other is likely to realize[29] — often even without any technical expertise. ("Wait, I never received that message!") Essentially, the courts are designed to preempted this kind of forgery through their adversarial nature.

The system is far from perfect. But we believe, at least, that the same principles that underlie the current court system's imperfect efforts to address forgery of evidence apply and generalize well to deniable communication.

5.1 Authentication

Authentication in court does not verify that a piece of evidence is true or "authentic" in the sense understood in colloquial speech or in computer science. It means that the party presenting the evidence must also "produce evidence sufficient [that a jury could reasonably find by a preponderance of the evidence] that the item is what the proponent claims it is.[30] A screenshot of a chat transcript, for example, could be sufficiently authenticated if the witness whose phone it is taken from testifies: "That is a transcript of a conversation between me and the defendant. We had that conversation in Signal on Friday at 9 P.M."[31]

This may seem incredibly insufficient for those accustomed to cryptographic definitions of authentication.[32] However, the goals of cryptographic authentication and authentication of evidence in court are different; the latter is not weaker, but incomparable in its goals. The authentication process for evidence is not designed to determine that a piece of evidence is certain to be what the party introducing it says it is. It is designed to allow the parties to introduce their evidence while creating an opportunity for their opponent to challenge it. Authentication via witness facilitates the assessment of whether evidence is credible based on the trustworthiness of the *witness*, not just the *evidence itself*. The witness is incentivized against lying because there are criminal penalties (for the crime of perjury) for doing so.[33] If the witness lies anyway, that may be indicated by other clues or contradictory evidence. This process works the same whether the evidence is a cryptographically deniable chat transcript, an unsigned letter, or a burnt piece of toast. The opposing party has a chance to cross-examine the witness in front of the jury. If there is reason to believe that the witness used deniability to alter the messages presented, they can ask the witness about it. The jury then decides how much bearing, if any, the transcript has on their verdict. Thus, the idea is that authentication of a deniable chat transcript via witness could be judged by the jury similarly to a witness describing a conversation held during a walk in the woods.

[26] Fed. R. Evid. 901(a). Here, *authentication* is a legal term of art, whose meaning differs completely from the cryptographic term of art in Section 3.
[27] Fed. R. Evid. 104(b).
[28] We are not aware of specific instances where forgeries of secure communication transcripts have come up as evidence in courts. We would be interested to hear of any examples. The closest we have come across is an interesting (as far as we know) hypothetical concern expressed by a Kenyan attorney that the blue "double checkmark" indicating that a WhatsApp message has been received and read might be misused as a "proof" in court that a recipient read a message, even if that message was deliberately deleted by the sender immediately afterward to prevent the recipient from reading it [31, §3.2.1].
[29] An opposing party might not be able to identify a discrepancy if the evidence was in the sole custody of their opponent, such as a business record. But then deniability is not needed for a forgery, so deniability still does not introduce any novel challenges.
[30] Fed. R. Evid. 901(a).
[31] *See* Fed. R. Evid. 901(b)(1).
[32] When we write "authentication," we mean authentication of evidence, not cryptographic authentication, unless otherwise specified.
[33] See, for example, 18 U.S.C. § 1621 (1994).

5.2 Relevance

A piece of evidence is considered *relevant* when it "has a tendency to make a fact more or less probable than it would be without the evidence"[34] and the fact is "of consequence" to the case.[35] "Tendency to make a fact more or less probable" is a flexible and inclusive phrasing that tries to get at the question of whether a person might reasonably assess a fact's likelihood of being true differently if they didn't have the evidence. The system is designed to put as much potentially useful evidence in front of the jury as possible.

From a pure mathematical understanding of deniability, introducing a transcript from a communications service which implements deniability would not seem to have a tendency to make any fact more or less probable. After all, such a transcript is not cryptographic evidence that any party to the communication actually sent the pictured messages. What is more, within the cryptographic model, the deniability of the communication scheme may mean that *any* proffered message is as likely as any other one to be the real one sent. By that same reasoning, introducing verbal testimony from a witness to a crime would not seem to have a tendency to make any fact more or less probable, because people can lie. However, when taking into account additional contextual factors, such as those introduced in our credibility framework, both the verbal testimony and the WhatsApp transcript could reasonably influence a juror's assessment of what likely happened — and accordingly, both would generally meet the relevance standard.

The rules of evidence were written to determine the relevance of evidence before cryptographic verification existed, and aim to handle, as fairly as possible, the scenarios in which it is most difficult to determine original from forgery. Cryptographic verification is not irrelevant to this process, and can streamline it when applicable — but the broader process still exists to handle the vast majority of evidentiary situations where cryptographic verification is not determinative, including situations involving deniable communication. As there will almost always be factors outside of cryptographic models that indicate whether a transcript is the original or not, even judges who deeply understood cryptographic deniability would be justified in determining that a deniable chat transcript, in the context of broader circumstances and other evidence, could have at least a *tendency* to make a fact more or less probable.

5.3 Expert testimony

Where a technical or scientific concept that is not widely understood is relevant to the case, courts also have mechanisms regarding *expert testimony* to explain that concept. Thus, a cryptographer by trade or education would be able to testify as to the effects of deniability on the veracity of chat transcripts. However, it is worth noting that cases often suffer from "dueling experts" where both sides call expert witnesses with relevant qualifications who present apparently opposite information, even though there is only one side well supported by scientific evidence [46]. Because judges and juries are not experts in the topics covered by most of the cases they try, they may side with the expert presenting bad evidence [46]. Law enforcement agencies have a history of introducing dubious forensic evidence, including ballistic and fingerprint analysis, that is scientifically unsound, sometimes while attacking defendants' scientific experts inside and outside of the courtroom [18]. Similarly, for decades the tobacco industry avoided liability by presenting experts who dismissed the epidemiologic association between smoking and lung cancer because no one knew how cells actually became cancerous [20]. Thus, bringing in experts is no panacea; it is important to consider the possibility that if a party tried to combat the presentation of forged deniable transcripts by presenting expert testimony, their opposing party might find an expert who would find some way to argue that the transcript was in fact the original, even though

[34]This phrasing may rankle the frequentists among us. Indeed, the Federal Rules of Evidence implicitly endorse Bayesian over frequentist probabilistic reasoning. We believe this is the right choice for the context; a full discussion is beyond the scope of this paper.
[35]Fed. R. Evid 401.

deniability is a settled cryptographic method. (To our knowledge, no such example has yet occurred.) Simply pointing out a discrepancy in a transcript without presenting expert testimony may sometimes be the best way to call the jury's attention to a possible forgery, without starting a battle of experts. That said, in some situations, such as where forgery by a sophisticated actor leveraging cryptographic deniability is suspected, bringing in an expert may be necessary to thoroughly explain the situation in the courtroom.

5.4 Example

To walk through the introduction of evidence at trial, consider a basic example: Prosecutor Pat is litigating the trial of Defendant Dan. Pat is claiming that Dan stole Witness Wendy's wallet and admitted to it in a handwritten ransom letter. Wendy approached Pat with the letter, shown in Figure 2, which requests $100 in exchange for the wallet. Pat wants to introduce that letter as evidence at trial.

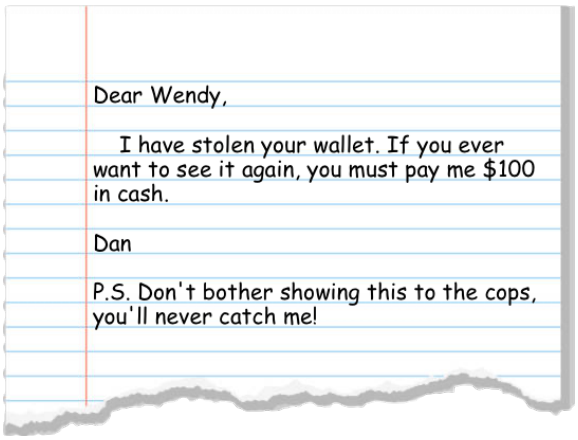

Fig. 2. Dan's ransom note to Wendy

To introduce the letter, Pat has to show it to Dan's lawyer, authenticate it through a witness, in this case, Wendy, and ask the judge to enter it into evidence. To block Pat from introducing the evidence, Dan's lawyer can argue that it is insufficiently authenticated, or that it is not relevant and/or admissible. Unless Wendy comes off as clearly lying, the judge will allow Pat to present it to the jury. The jury will then have to decide how credible they think the letter is.

Now, imagine instead of giving Pat a handwritten letter, Wendy shows Pat a transcript from Signal. To introduce the transcript at trial, Pat will follow the exact same steps as the letter. Dan's lawyer can still argue that it is insufficiently authenticated, or that it is not relevant and/or admissible. And again, unless Wendy comes off as clearly lying, the judge will allow Pat to present it to the jury, which will decide how credible they think the transcript is.

There are few scenarios where it would be worth the time of Dan's lawyers to argue against the introduction of the Signal transcript. The most likely would be if Pat tried to introduce a transcript copied from Wendy's phone that did not match the transcript on Dan's phone. It is important to note that this discrepancy would be immediately apparent to Dan's lawyers and easily impeachable even without technical knowledge of deniability. The mere fact of the discrepancy would enable Dan's lawyers to argue against the introduction of the evidence. While they could introduce technical experts to explain deniability, even without expert testimony, the jury's consideration of the evidence would be similar to consideration of conflicting testimony about a walk in the woods.

Even if expert testimony led to the dismissal of the evidence, it is unlikely to be instrumental in acquitting Dan. It is implausible that a case with only a chat transcript as evidence would make it to trial. While transcripts could theoretically be pivotal in securing a conviction, it is more common for an investigation to begin with law enforcement receiving a communication and only lead to charges once it finds more extensive evidence.

6 DISCUSSION

Existing legal systems are deeply imperfect. This paper is not intended to extol courts or evidentiary systems. Our points are conditioned on a legal system existing in its current form (and we believe many of our points are likely to generalize to a range of other legal systems). We argue that legal systems generally handle deniable communication evidence as their current evidentiary rules intend, rather than mishandling it due to a lack of technical expertise; and moreover, that those evidentiary rules have been developed with a detailed and explicit consideration of exactly the kinds of issues that cryptographic deniability raises, over a long time. As such, the legal system does not require a paradigm shift to deal with deniability; rather, to the extent that the legal system does not deal with deniability satisfactorily, deniability is an additional example of the problems that the system has grappled with and will continue to face.

Judges' and laypeople's understanding of communication technologies does matter. Although, as we have discussed, all types of evidence are in some sense forgeable, judges' and jurors' perception of the fact and ease of forgeability varies between types of evidence and can have a significant bearing on the outcome of a case. For example, when hearing oral testimony, a jury will intuitively consider the fact that people can lie, but they may not as readily think of forgeability when weighing a chat transcript. A jury will also, in most cases, have an intuitive way of evaluating whether a witness is lying but likely will not have the same skills to evaluate digital evidence [40]. As noted above, the evidentiary system offers some "bootstrapping" of this gap in knowledge by tying the authenticity of evidence back to oral testimony.

The intuition of the average person, and therefore the average juror, regarding forgeability naturally changes over time. When photo editing software was first introduced, similar concerns were raised about whether juries could trust photographic evidence. Now, we expect that many jurors would intuitively understand that photographs can be manipulated (even if not exactly how), and to evaluate evidence with that knowledge even without prompting. Deepfakes, discussed in the following paragraph, are the newest technology to present similar challenges. Explaining the difficulty and likelihood of forgery to jurors in an intuitive way can be outcome determinative in court cases, so it is important for the administration of justice to better understand how people receive and interpret this kind of instruction. The studies discussed in Section 2 are important initial steps in this process.

Deepfakes raise interesting new issues. Deepfakes make the creation of believable forged images far easier than prior technologies. Although the process of creating a deepfake is technically sophisticated, access to forgeries is quite easy at low cost for laypeople because, at least until recently, online platforms incentivized this access [28]. They also impact the threshold of believability — because of media hype and political relevance, the public is hyper-aware of the existence of deepfakes even if they do not understand the underlying technology. In the courtroom, generally understanding how the technology works may be relevant to prevent the most absurd uses (such as [17]). This is likely to present significant challenges in the future, as the technology is complex and difficult for laypeople to understand. Perhaps courtrooms are not the most important venue for this concern. Because of the high level of scrutiny and the likelihood that forensics will be applied to evidence in a legal case, a deepfake presented as an authentic image would be fairly

likely to be discovered. In any case, we expect deepfakes to raise high profile issues in courtrooms in the coming years, and anticipate further public discussion on this topic with interest.[36]

7 CONCLUSION

The value of deniability lies largely outside the courts. But deniability can help provide users with communications secured against a variety of real threats, which support freedom of expression in practical scenarios. Credibility is an essential consideration for implementing deniability features in practice, and plays a major and previously underacknowledged role in the utility of deniable communications. Since credibility is inherently sociotechnical, we believe there is a great breadth of possible systems which could integrate this concept to provide novel value.

ACKNOWLEDGMENTS

We are grateful to Alastair Beresford, Joseph Bonneau, Ran Canetti, James Grimmelmann, and Alice Hutchings for helpful discussions and feedback.

REFERENCES

[1] A. Acquisti and J. Grossklags. Privacy and rationality in individual decision making. *IEEE Secur. Priv.*, 3(1):26–33, 2005. DOI: 10.1109/MSP.2005.22. URL: https://doi.org/10.1109/MSP.2005.22.

[2] S. Ahmad, S. Rass, and P. Schartner. False-bottom encryption: deniable encryption from secret sharing. *IEEE Access*, 11, 2023.

[3] American Academy of Forensic Sciences. What is forensic science? URL: https://www.aafs.org/careers-forensic-science/what-forensic-science. PERMALINK: https://web.archive.org/web/20250730015157/https://www.aafs.org/careers-forensic-science/what-forensic-science.

[4] J. Bartusek, V. Goyal, D. Khurana, G. Malavolta, J. Raizes, and B. Roberts. Software with certified deletion. In M. Joye and G. Leander, editors, *Advances in Cryptology – EUROCRYPT 2024*, pages 85–111. Springer Nature Switzerland, 2024. ISBN: 978-3-031-58737-5. DOI: https://doi.org/10.1007/978-3-031-58737-5_4.

[5] M. Bateson, L. Callow, J. R. Holmes, M. L. Redmond Roche, and D. Nettle. Do images of 'watching eyes' induce behaviour that is more pro-social or more normative? a field experiment on littering. *PLOS ONE*, 2013. DOI: 10.1371/journal.pone.0082055. URL: https://doi.org/10.1371/journal.pone.0082055.

[6] R. Canetti, C. Dwork, M. Naor, and R. Ostrovsky. Deniable encryption. Cryptology ePrint Archive, Paper 1996/002, 1996. URL: https://eprint.iacr.org/1996/002. PERMALINK: https://perma.cc/D5WL-4JNJ.

[7] R. Canetti, S. Park, and O. Poburinnaya. Fully deniable interactive encryption. Cryptology ePrint Archive, Paper 2018/1244, 2018. URL: https://eprint.iacr.org/2018/1244. PERMALINK: https://perma.cc/JZV5-MWUL.

[8] D. J. Capra and L. L. Richter. Long live the federal rules of evidence! *George Mason Law Review*, 2024. URL: https://lawreview.gmu.edu/wp-content/uploads/2023/12/Capra-Richter-31-Geo-Mason-L-Rev-1-2024.pdf. PERMALINK: https://perma.cc/EQ4F-HVRC.

[9] E. Casey. *Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet*. Academic Press, Inc., 3rd edition, 2011. ISBN: 0123742684.

[36] For example, President Trump has suggested that, in the future, "If something happens that's really bad, maybe I'll have to just blame AI." [38]

[10] S. Chakraborty, D. Hofheinz, U. Maurer, and G. Rito. Deniable authentication when signing keys leak. In C. Hazay and M. Stam, editors, *Advances in Cryptology – EUROCRYPT 2023*, pages 69–100. Springer Nature Switzerland, 2023. ISBN: 978-3-031-30620-4. DOI: https://doi.org/10.1007/978-3-031-30620-4_3.

[11] A. Cohen and S. Park. Compelled decryption and the fifth amendment: exploring the technical boundaries. *Harvard Journal of Law & Technology*, 2018. URL: https://jolt.law.harvard.edu/assets/articlePDFs/v32/32HarvJLTech169.pdf. PERMALINK: https://perma.cc/RKT2-CX9G.

[12] D. Collins, S. Colombo, and L. Huguenin-Dumittan. Real-world deniability in messaging. Cryptology ePrint Archive, Paper 2023/403, 2023. URL: https://eprint.iacr.org/2023/403. PERMALINK: https://perma.cc/NFW2-VTDG.

[13] K. Eichhorn. Why an internet that nevers forgets is especially bad for young people, 2019. URL: https://www.technologyreview.com/2019/12/27/131123/internet-that-never-forgets-bad-for-young-people-online-permanence/. PERMALINK: https://perma.cc/U4CP-PEDL.

[14] J. Fairoze. *On Zeroization in Secure Messaging*. Master's thesis, University of Melbourne, 2021. Unpublished.

[15] N. E. Frye and M. M. Dornisch. When is trust not enough? the role of perceived privacy of communication tools in comfort with self-disclosure. *Computers in Human Behavior*, 2010. DOI: https://doi.org/10.1016/j.chb.2010.03.016. URL: https://www.sciencedirect.com/science/article/pii/S0747563210000567.

[16] N. E. Frye and M. M. Dornisch. When is trust not enough? the role of perceived privacy of communication tools in comfort with self-disclosure. *Computers in Human Behavior*, 26(5):1120–1127, 2010. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2010.03.016. URL: https://www.sciencedirect.com/science/article/pii/S0747563210000567.

[17] M. Gault and J. Koebler. 'i loved that ai:' judge moved by ai-generated avatar of man killed in road rage incident. 404 Media, 2025. URL: https://www.404media.co/i-loved-that-ai-judge-moved-by-ai-generated-avatar-of-man-killed-in-road-rage-incident/. PERMALINK: https://perma.cc/6SX9-S46X.

[18] P. C. Giannelli. Daubert and forensic science: the pitfalls of law enforcement control of scientific research. *University of Illinois Law Review*, 2011. URL: https://illinoislawreview.org/wp-content/ilr-content/articles/2011/1/Giannelli.pdf. PERMALINK: https://perma.cc/U9P4-HQLH.

[19] D. K. Gillmor. Disappearing messages don't work — and they're great. ACLU News & Commentary, 2025. URL: https://www.aclu.org/news/privacy-technology/disappearing-messages-dont-work-and-theyre-great. PERMALINK: https://perma.cc/FB9U-F5SN.

[20] S. C. Gold. A fitting vision of science for the courtroom. *Wake Forest Journal of Law and Policy*, 2013. URL: https://ssrn.com/abstract=2101454.

[21] S. E. Igo. *The Known Citizen*. Harvard University Press, 2018.

[22] International Association for Cryptologic Research. Cryptology ePrint Archive. URL: https://eprint.iacr.org.

[23] D. Kaye. Report on encryption, anonymity, and the human rights framework. https://www.undocs.org/A/HRC/29/32, 2015. PERMALINK: https://perma.cc/QN7A-R9V5.

[24] J. Li, Z. Lin, J. Caballero, Y. Zhang, and D. Gu. K-hunt: pinpointing insecure cryptographic keys from execution traces. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, 412–425. Association for Computing Machinery, 2018. ISBN: 9781450356930. DOI: 10.1145/3243734.3243783. URL: https://doi.org/10.1145/3243734.3243783.

[25] M. H. Ligh, A. Case, J. Levy, and A. Walters. *The Art of Memory Forensics: Detecting Malware and Threats in Windows, Linux, and Mac Memory*. Wiley, 1st edition, 2014.

[26] K. Lisiecka, A. Rychwalska, K. Samson, K. Łucznik, M. Ziembowicz, A. Szóstek, and A. Nowak. Medium moderates the message. how users adjust their communication trajectories to different media in collaborative task solving. *PLOS ONE*, 11, June 2016. DOI: 10.1371/journal.pone.0157827. URL: https://doi.org/10.1371/journal.pone.0157827.

[27] K. Litman-Navarro. We read 150 privacy policies. They were an incomprehensible disaster. https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html, 2019. PERMALINK: https://perma.cc/5XJJ-4CQV.

[28] E. Maiberg and S. Cole. Mr. deepfakes, the biggest deepfake porn site on the internet, says it's shutting down for good, 2025. URL: https://www.404media.co/mr-deepfakes-the-biggest-deepfake-porn-site-on-the-internet-says-its-shutting-down-for-good/. PERMALINK: https://perma.cc/GE6L-DG2K.

[29] M. Marlinspike. The x3dh key agreement protocol, 2016. URL: https://signal.org/docs/specifications/x3dh. PERMALINK: https://perma.cc/LCV5-KW9T.

[30] A. M. McDonald and L. F. Cranor. The cost of reading privacy policies. *I/S: A Journal of Law & Policy for the Information Society*, 4:543–897, 2009.

[31] C. W. Munyendo, K. Owens, F. Strong, S. Wang, A. J. Aviv, T. Kohno, and F. Roesner. "you have to ignore the dangers": user perceptions of the security and privacy benefits of whatsapp mods. In M. Blanton, W. Enck, and C. Nita-Rotaru, editors, *IEEE Symposium on Security and Privacy, SP 2025, San Francisco, CA, USA, May 12-15, 2025*, pages 4515–4533. IEEE, 2025. DOI: 10.1109/SP61157.2025.00087. URL: https://doi.org/10.1109/SP61157.2025.00087.

[32] National Institute of Standards and Technology. What is forensic science? URL: https://www.nist.gov/forensic-science. PERMALINK: https://perma.cc/4AUR-NUNL.

[33] H. Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79, 1, 2004.

[34] D. Oba and J. Berger. How communication mediums shape the message. *Journal of Consumer Psychology*, 34(3), 2024. DOI: https://doi.org/10.1002/jcpy.1372. URL: https://myscp.onlinelibrary.wiley.com/doi/abs/10.1002/jcpy.1372.

[35] T. A. H. D. of the English Language. Deniable. URL: https://ahdictionary.com/word/search.html?q=deniable. PERMALINK: https://perma.cc/AQD9-QQRC.

[36] T. A. H. D. of the English Language. Forgery. URL: https://ahdictionary.com/word/search.html?q=forgery. PERMALINK: https://perma.cc/XQK2-LWQX.

[37] H. L. Packer. A tale of two typewriters. *Stanford Law Review*, 1958. URL: https://doi.org/10.2307/1226822.

[38] M. L. Price. Trump says video showing items thrown from white house is ai after his team indicates it's real. AP News, 2025. URL: https://apnews.com/article/trump-white-house-window-ai-9dad119bb6de1519582db036dc0726d7. PERMALINK: https://web.archive.org/web/20250904144114/https://apnews.com/article/trump-white-house-window-ai-9dad119bb6de1519582db036dc0726d7.

[39] B. Prince. The mystery of the truecrypt encryption software shutdown. Dark Reading, 2014. URL: https://www.darkreading.com/endpoint-security/the-mystery-of-the-truecrypt-encryption-software-shutdown. PERMALINK: https://web.archive.org/web/20240910142704/https://www.darkreading.com/endpoint-security/the-mystery-of-the-truecrypt-encryption-software-shutdown.

[40] A. Rajendran, T. K. Yadav, M. Al-Jbour, F. M. Mares Solano, K. Seamons, and J. Reynolds. Deniable encrypted messaging: user understanding after hands-on social experience. In *Proceedings of the 2024 European Symposium on Usable Security*, 155–171. Association for Computing Machinery, 2024. ISBN: 9798400717963. DOI: 10.1145/3688459.3688479. URL: https://doi.org/10.1145/3688459.3688479.

[41] N. Reitinger, N. Malkin, O. Akgul, M. L. Mazurek, and I. Miers. Is cryptographic deniability sufficient*f* non-expert perceptions of deniability in secure messaging. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 274–292, 2023. DOI: 10.1109/SP46215.2023.10179361.

[42] D. J. Solove. The myth of the privacy paradox. *George Washington Law Review*, 89, 1, 2021.

[43] M. Van Kleek, I. Liccardi, R. Binns, J. Zhao, D. J. Weitzner, and N. Shadbolt. Better the devil you know: exposing the data sharing practices of smartphone apps. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.

[44] N. Vatandas, R. Gennaro, B. Ithurburn, and H. Krawczyk. On the cryptographic deniability of the signal protocol. Cryptology ePrint Archive, Paper 2021/642, 2021. URL: https://eprint.iacr.org/2021/642. PERMALINK: https://perma.cc/UY62-9BZJ.

[45] T. K. Yadav, D. Gosain, and K. Seamons. Cryptographic deniability: a multi-perspective study of user perceptions and expectations. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 3637–3654. USENIX Association, 2023. ISBN: 978-1-939133-37-3. URL: https://www.usenix.org/conference/usenixsecurity23/presentation/yadav. PERMALINK: https://perma.cc/PDM5-LMSW.

[46] K. K. Yee. Dueling experts and imperfect verification. *International Review of Law & Economics*, 2008. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1312222.