CONTENT MODERATION ON END-TO-END ENCRYPTED SYSTEMS: A LEGAL ANALYSIS

Charles Duan James Grimmelmann

End-to-end encrypted online platforms are increasingly common in the digital ecosystem, found both in dedicated apps like Signal and widely adopted platforms like Android Messages. Though such encryption protects privacy and advances human rights, the law enforcement community and others have raised criticisms that end-to-end encryption shields bad behavior, preventing the platforms or government authorities from intercepting and responding to criminal activity, child exploitation, malware scams, and disinformation campaigns. At a time when major Internet platforms are under scrutiny for content moderation practices, the question of whether end-to-end encryption interferes with effective content moderation is of serious concern.

Computer science researchers have responded to this challenge with a suite of technologies that enable content moderation on end-to-end encrypted platforms. Are these new technologies legal? This Article analyzes these new technologies in light of several major federal communication privacy regimes: the Wiretap Act, the Stored Communications Act, and the Communications Assistance for Law Enforcement Act.

While generally we find that these content moderation technologies would pass muster under these statutes, the answers are not as clear-cut as one might hope. The advanced cryptographic techniques that these new content moderation strategies employ raise multiple unsettled questions of law under the communication privacy regimes considered. This legal uncertainty arises not because of the ambiguous ethical nature of the technologies themselves, but because the decades-old statutes failed to accommodate, or indeed contemplate, the innovations in cryptography that enable content moderation to coexist with encryption. To the extent that platforms are limited in their ability to moderate end-to-end encrypted content, then, those limits may arise not from the technology but from the law.

Introduction

Encryption has costs. Most obviously, there are technical costs. Encrypting and decrypting messages takes time and computing power, and creating secure encrypted systems takes immense engineering effort. Most controversially, there are policy costs. Law enforcement groups object when encryption works as intended, because it makes it harder for authorities to read suspects' communications. And most subtly, there are safety costs. Encrypting messages makes it harder to protect users.

On modern communications platforms, content moderation plays a central role in keeping users safe from spam, harassment, and abuse.[1] To moderate content, a platform must know what that content is. But when content is encrypted so that not even the platform itself can read it – when it is protected with *end-to-end encryption*, or E2EE for short – standard techniques of content moderation become impossible. Humans cannot read the messages to see if they contain threats of violence; computers cannot scan them to see if they contain child sexual abuse material (CSAM).

In short, it has appeared that encryption and moderation are incompatible. Heightening the irony, encryption itself is a also safety technology,

1. *See* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 Harv. L. Rev. 1598 (2017).

because privacy is a form of safety.[2] Encryption advances interests such as data security, privacy, free speech, free association, and other constitutional and human rights.[3] Weakening encryption to enable third-party content access (either for content moderation or law enforcement), besides undermining these benefits, creates potentially risky security vulnerabilities.[4] So the fact that E2EE enhances one form of safety (privacy) while undermining another (protection from abuse) seems like a tragic but inevitable tradeoff.

In the last few years, however, computer-science researchers have shown that encryption and moderation are not so incompatible after all. Research teams around the world have developed ways to support content moderation and abuse prevention that do not require letting a communications platform view the unencrypted contents of messages. Indeed, this is such a fruitful area of research that there are already works in the technical literature that taxonomize and systematize research on so-called "content moderation on end-to-end encrypted systems." Some of the interest driv-

---

2. A. Michael Froomkin & Zak Colangelo, *Privacy as Safety*, 95 Wash. L. Rev. 141 (2020).

3. *See, e.g.,* Orin S. Kerr, *The Fourth Amendment in Cyberspace: Can Encryption Create a Reasonable Expectation of Privacy*, 33 Conn. L. Rev. 503, 503–04 (2000–2001); A. Michael Froomkin, *Metaphor Is the Key: Cryptography, the Clipper Chip, and the Constitution*, 143 U. Pa. L. Rev. 709, 810–43 (1994–1995); Christopher Soghoian, *Caught in the Cloud: Privacy, Encryption, and Government Back Doors in the Web 2.0 Era*, 8 J. on Telecomms. & High Tech. L. 359, 375–76, 398–99 (2010); Jan H. Samoriski et al., *Encryption and the First Amendment*, 2 Commc'n L. & Pol'y 417 (1997); Geoffrey Gordon, Note, *Breaking the Code: What Encryption Means for the First Amendment and Human Rights*, 32 Colum. Hum. Rts. L. Rev. 477 (2001); Gwynne B. Barrett, Note, *Law of Diminishing Privacy Rights: Encryption Escrow and the Dilution of Associational Freedoms in Cyberspace*, 15 N.Y.L. Sch. J. Hum. Rts. 115 (1998); Sean J. Edgett, *Double-Clicking on Fourth Amendment Protection: Encryption Creates a Reasonable Expectation of Privacy*, 30 Pepp. L. Rev. 339 (2002–2003).

4. *See, e.g.,* Richard A. Clarke et al., Liberty and Security in a Changing World: Report and Recommendations of The President's Review Group on Intelligence and Communications Technologies 216–19 (2013), https://obamawhitehouse.archives.gov/sites/default/files/docs/2013-12-12_rg_final_report.pdf (recommending that the United States "not in any way subvert, undermine, weaken, or make vulnerable generally available commercial software" for encryption); Hal Abelson et al., The Risks of Key Recovery, Key Escrow, and Trusted Third-Party Encryption 10–13 (1997), https://academiccommons.columbia.edu/doi/10.7916/D8GM8F2W; Peter Swire & Kenesa Ahmad, *Encryption and Globalization*, 13 Colum. Sci. & Tech. L. Rev. 416, 432–33 (2012).

ing this research comes from platforms themselves,[5] while others are motivated by the search for technical tools to help address policy problems.[6]

One of these techniques, known in the literature as "message franking," allows the recipient of an abusive message to report it to a moderator, *with receipts*.[7] By virtue of clever construction of the receipts, message franking guarantees that recipients can prove that the message they are reporting is authentic, that recipients cannot submit false reports, and that no one besides the recipient of a message can report a message—or even learn anything about it. An extension of message franking, called "forward tracing," allows the platform to trace a reported message back to its original sender, even if it has been forwarded repeatedly—but again without compromising the privacy of messages that are not reported.[8]

Another broad class of new techniques involve automated scanning of messages to detect problematic content, such as unsolicited photographs of genitalia.[9] Again, it seems like a paradox: how can a platform scan encrypted content? Here, the answer lies in a class of algorithms called "homomorphic encryption."[10] The idea behind homomorphic encryption is that an untrusted party can perform computations on content without knowing what the content is. Imagine a blindfolded chef wearing thick mittens, who follows instructions to take things out of a box, chop them up, put them in the oven for an hour at 350 degrees, and then put it back in the

5. Will Cathcart, *Encryption Has Never Been More Essential—Or Threatened*, WIRED (Apr. 5, 2021), https://www.wired.com/story/opinion-encryption-has-never-been-more-essential-or-threatened/ (op-ed by head of WhatsApp) ("[B]y employing sophisticated techniques to analyze metadata, user reports, and other unencrypted information, we ban millions of dangerous accounts every year.").

6. *Everything in Moderation?*, HORIZON DIGIT. ECON. RSCH. (June 6, 2022), https://www.horizon.ac.uk/everything-in-moderation/ ("E2E encryption presents challenges in dealing with misinformation, disinformation, potentially harmful or illegal content, and striking a balance with freedom of speech."); Ariadna Matamoros-Fernández, *Encryption Poses Distinct New Problems: The Case of WhatsApp* 9, *in* 9 INTERNET POL'Y REV. 7 (2020), https://policyreview.info/pdf/policyreview-2020-4-1512.pdf ("The pervasiveness of encrypted platforms in mediating everyday life in some parts of the world is a reminder that viable content moderation measures without breaking encryption are needed.").

7. *See infra* Section III.A.

8. *See infra* Section III.B.

9. *See infra* Section III.C.

10. *See infra* pp. 56–57.

box. This chef can roast vegetables for you, but doesn't learn whether you were roasting potatoes or parsnips. Similarly, homomorphic encryption allows a platform to run a bad-content detector on a message, *and report the result to the recipient*, without the platform itself learning anything about the content or the result.[11]

Despite the recent explosion in research on moderation in the presence of encryption, legal scholarship has given almost no attention to these new technologies.[12] A search of HeinOnline as of January 31, 2023 for "message franking" returns no results. Only a small handful of position papers have discussed their implications for law and policy.[13]

This omission is unfortunate, because the assumption that encryption and moderation are incompossible has been baked into the long-running debates about whether and how to regulate the use of E2EE. Some commentators, for example, have criticized platforms' decisions to deploy E2EE by decrying the consequences for reduced content moderation.[14]

---

11. *See* David Wong, Real-World Cryptography § 15.2, fig.15.6 (2021).

12. An exception is client-side scanning, since that technology has received substantial attention in the press. *See* Timothy Gernand, *Scanning iPhones to Save Children: Apple's on-Device Hashing Algorithm Should Survive a Fourth Amendment Challenge*, 127 Dick. L. Rev. 307, 319–20 (2022); Nicholas A. Weigel, *Apple's "Communication Safety" Feature for Child Users: Implications for Law Enforcement's Ability to Compel iMessage Decryption*, 25 Stan. Tech. L. Rev. 210, 216–17 (2022). Nevertheless, the client-side scanning protocols in the computer science literature go far beyond the relatively simple image-flagging proposals that the legal scholarship has considered. *See infra* Section III.D.

13. *See, e.g.*, Ian Levy & Crispin Robinson, Thoughts on Child Safety on Commodity Platforms (July 21, 2022) (unpublished manuscript), https://arxiv.org/abs/2207.09506; Hal Abelson et al., Bugs in Our Pockets: The Risks of Client-Side Scanning (Oct. 15, 2021) (unpublished manuscript), https://arxiv.org/abs/2110.07450; Jonathan Mayer, Content Moderation for End-to-End Encrypted Messaging (Oct. 6, 2019) (unpublished discussion paper), https://www.cs.princeton.edu/~jrmayer/papers/Content_Moderation_for_End-to-End_Encrypted_Messaging.pdf; Bus. for Soc. Resp., Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption (2022), https://www.bsr.org/reports/bsr-meta-human-rights-impact-assessment-e2ee-report.pdf; *see also* Gurshabad Grover et al., *The Ministry and the Trace: Subverting End-to-End Encryption* 16, *in* 14 NUJS L. Rev. (2021), http://nujslawreview.org/2021/07/09/the-ministry-and-the-trace-subverting-end-to-end-encryption/ (considering legality of forward tracing in view of India's constitutional right to privacy).

14. *See* Siva Vaidhyanathan, *Be Careful Taking Sides in Mark Zuckerberg vs. William Barr*, Slate (Oct. 4, 2019), https://slate.com/technology/2019/10/facebook-encryption-

This Article aims to bridge this gap between the computer-science and legal literatures. It makes three contributions.

First, we introduce and explain the current universe of content moderation technologies for end-to-end encrypted platforms. We focus particularly on technologies drawn specifically toward encrypted communications, disregarding already-existing techniques that work independent of encryption, such as metadata analysis or platform affordance modifications.[15] Some of the technologies we discuss work around the encryption, either by moderating content automatically on users' devices[16] or by manipulating the encrypted content in ways that have predictable effects on the content despite not revealing it.[17] Others take advantage of users to flag content for moderators' review. The thrust of these technologies is to provide the platform with certainty about who to take action against in response to a valid user report of abuse, and the challenge is providing that certainty without compromising the confidentiality of unreported messages.[18]

Second, we provide a careful legal analysis of these moderation technologies *vis à vis* the federal communication privacy laws. Application of these laws to the content moderation technologies in question, we find, is not always straightforward. While in many cases the statutory definitions map well onto the communicative aspects of the technologies, there are often ambiguities in the statutes and case law, leaving it unclear how a court

---

mark-zuckerberg-william-barr.html (suggesting that Facebook has "motivations to install encryption" so that "Facebook can't be held responsible for failing to keep its system free of calls for violence, harassment, or hate speech"); Natasha Lomas, *UK Tells Messaging Apps Not to Use E2E Encryption for Kids' Accounts*, TechCrunch (June 30, 2021), https://techcrunch.com/2021/06/30/uk-tells-messaging-apps-not-to-use-e2e-encryption-for-kids-accounts/ (noting U.K. government guidance to platforms that "[e]nd-to-end encryption makes it more difficult for you to identify illegal and harmful content occurring on private channels"). *But see* Mike Masnick, *The DOJ Is Conflating the Content Moderation Debate with the Encryption Debate: Don't Let Them*, Techdirt (Oct. 8, 2019), https://www.techdirt.com/2019/10/08/doj-is-conflating-content-moderation-debate-with-encryption-debate-dont-let-them/.

15. *See* Matamoros-Fernández, *supra* note 6, at 7–8. For example, to limit the rapid spread of disinformation, WhatsApp introduced limits on the number of times messages could be forwarded. *See id.* at 8.

16. *See infra* Section III.D.

17. *See infra* Section III.C.

18. *See infra* Section III.A; *infra* Section III.B.

would rule. In the worst case, a court might hold that these types of content moderation are actually *illegal*, perversely putting platforms and users back to the stark and unpleasant choice between encryption and moderation. We wish we could say that United States communications privacy laws clearly rule out such an outcome—but unfortunately they do not.

Third, we use our legal analysis as a basis to critique both the statutes and the technologies. Statutory ambiguities and unresolved questions of legal interpretation highlight potential areas for reform, to bring the communication privacy laws in line with developments in cryptographic research since those statutes were first enacted almost half a century ago. But those ambiguities and questions also highlight potential areas of ethical concern with the content moderation technologies themselves. After all, the communication privacy laws are intended to reflect, however accurately, intuitive and societal norms of privacy. Discrepancies between the statutes and the technologies are a starting point for a larger conversation on how those technologies impact privacy interests, an especially significant conversation given the immense privacy benefits of end-to-end encryption.

Part I of the Article gives a technical and historical background on encryption technologies and how they came to pose content-moderation challenges.[19] Part II is a legal overview. It describes the five most relevant statutes: the Wiretap Act, the Stored Communications Act ("SCA"), the Pen Register Act ("PRA"), the Computer Fraud and Abuse Act ("CFAA"), and the Communications Assistance for Law Enforcement Act ("CALEA").[20] Part III, the heart of the Article, reviews four principal techniques for content moderation in the presence of E2EE: (1) message franking, (2) forward tracing, and automated scanning with homomorphic encryption either on (3) the platform's servers, or (4) on the user's devices. For each technique, it gives a technical overview, and then analyzes how that technique fares under the various communication privacy statutes.[21] Part IV then steps back to extract broader lessons for legal scholars and policy makers.[22]

19. *See infra* Part I.
20. *See infra* Part II.
21. *See infra* Part III.
22. *See infra* Part IV.

I. Background

A. Encryption Technologies

This section provides a brief overview of the fundamentals of modern cryptography. The reader who is already familiar with the concepts of public-key encryption and hash functions should feel free to skip ahead to the next section.[23]

Encryption is a process that keeps information confidential by rendering it unintelligible to outsiders. For a simple example, consider ROT-13 encryption, in which every letter in a text is replaced with the letter that is thirteen away in the alphabet:

```
ABCDEFGHIJKLMNOPQRSTUVWXYZ
          ⇓
NOPQRSTUVWXYZABCDEFGJIHKLM
```

$$\texttt{"WE THE PEOPLE"} \xrightarrow{\text{Encrypt}_{ROT-13}} \texttt{"JR GUR CRBCYR"}$$

The unencrypted data (`"WE THE PEOPLE"`) is called the *plaintext*, and the encrypted data (`"JR GUR CRBCYR"`) is called the *ciphertext*.[24] To an unauthorized party who doesn't know how the data has been encrypted, the ciphertext should should appear to be random. But an authorized party who knows that it has been encrypted using ROT-13 can recover the plaintext by undoing the letter-by-letter replacement:

$$\texttt{"JR GUR CRBCYR"} \xrightarrow{\text{Decrypt}_{ROT-13}} \texttt{"WE THE PEOPLE"}$$

ROT-13 is not a very good encryption algorithm, because its security collapses as soon as an eavesdropper (typically called an *adversary* in the cryptography literature) learns what algorithm is being used. A better

23. For more thorough overviews of central concepts in cryptography, see generally (in ascending order of detail) James Grimmelmann, Internet Law: Cases and Problems 40–45 (12th ed. 2022); Nat'l Acad. of Scis., Cryptography and the Intelligence Community: The Future of Encryption 16–34 (2022), https://nap.nationalacademies.org/read/26168; Mike Rosulek, The Joy of Cryptography (2021), https://joyofcryptography.com.

24. *See* Rosulek, *supra* note 23, at 10.

approach, and the one which is universally used today, is to use an algorithm that combines the plaintext with an additional piece of information, called an encryption *key*, to produce the ciphertext.[25] That way, even if the algorithm is publicly known, data encrypted with that algorithm will be indiscernible to an adversary so long as the key is kept appropriately secret.[26]

ROT-13 is weak because it has no separate key, but a closely related encryption algorithm is stronger because it does. In a "Caesar cipher," each letter in the alphabet is replaced with the letter that is k letters ahead of it in the alphabet (wrapping around at the end, so that `A` follows `Z`).[27] The number k is the key for a Caesar cipher; it can have any value from 1 to 26. The substitution for a Caesar cipher with a key of 1 is

```
ABCDEFGHIJKLMNOPQRSTUVWXYZ
                ⇓
BCDEFGHIJKLMNOPQRSTUVWXYZA
```

the Caesar cipher with a key of 2 is

```
ABCDEFGHIJKLMNOPQRSTUVWXYZ
                ⇓
CDEFGHIJKLMNOPQRSTUVWXYZAB
```

and so on. ROT-13 is just a Caesar cipher with a hardwired key of 13.

Caesar ciphers are a form of *symmetric-key* encryption because the key used to decrypt an encrypted ciphertext is the same as or easily derivable from the encryption key.[28] To encrypt a message using a Caesar cipher, shift every letter in the plaintext *forward* by k letters. To decrypt the message, shift every letter in the ciphertext *back* by k letters. Symmetric-key encryption can be simple, fast, and convenient, but it also has some significant disadvantages. One of them is the problem of key distribution. Every

25. *See id.* at 11.

26. *See* Orin S. Kerr & Bruce Schneier, *Encryption Workarounds*, 106 Geo. L.J. 989, 993 (2018) (describing Kerchoffs's Principle, specifying that "[a]n encryption algorithm should be secure if everything is known about it except the key").

27. *See* Dennis Luciano & Gordon Prichett, *Cryptology: From Caesar Ciphers to Public-Key Cryptosystems*, Coll. Mathematics J. 3–4 (1987), https://www.tandfonline.com/doi/abs/10.1080/07468342.1987.11973000.

28. *See* Rosulek, *supra* note 23, at 260.

pair of people who wish to communicate must coordinate in advance to agree on a secret key to use—and make sure that they don't accidentally reveal the key to an adversary in the process.[29]

To overcome this, *asymmetric* or *public-key* encryption ciphers use a pair of keys, called the *public key* and the *private key*.[30] The sender encrypts the message using the public key; the receiver decrypts the message using the private key.

$$\text{Plaintext} \xrightarrow{\text{Encrypt}_{\text{Public Key}}} \text{Ciphertext} \xrightarrow{\text{Decrypt}_{\text{Private Key}}} \text{Plaintext}$$

As the names suggest, in many widely used encryption schemes, the public key is truly public and the private key is truly private. A person who wants to receive messages will generate a key pair and widely distribute the public key so that anyone can use it to encrypt messages to them. But the person will keep the private key to themselves, so that they are the only person who can decrypt those messages.[31]

This asymmetry is what makes end-to-end encryption possible. Suppose that Alice and Bob want to exchange a message on a platform they don't trust. Alice encrypts the message using Bob's public key, and then hands the message off to the platform to deliver to Bob. When Bob receives Alice's message, he can decrypt it using his private key. But because Bob has not shared his private key with anyone, not even the platform cannot decrypt Alice's message.

By contrast, in a non-end-to-end encrypted messaging system, the platform has access to the plaintext. This does not mean there is no encryption involved. The communication from Alice to the platform might be encrypted, and so might the communication from the platform to Bob. And the platform might encrypt the message when it is "at rest" in storage, to keep it safe against hackers. The crucial point, however, is that any encryption that takes place involves a decryption key the platform has access to. It can use that key whenever it wants, so Alice and Bob must trust the platform not to misuse that power (or be compelled to misuse it). But when

29. *See id.* at 12 (noting difficulty of "key distribution").

30. *See id.* at 260.

31. Crucially, it is not feasible to derive the private key from the public key. This is what distinguishes asymmetric encryption from symmetric encryption like a Caesar cipher, where the decryption key is just 26 minus the encryption key.

Alice encrypts the message to Bob herself using Bob's public key, Alice and Bob only need to trust each other, and not also the platform.

Public-key encryption is surprisingly versatile. In addition to protecting the privacy of a message, it can be used to establish that the message is authentic. Abstractly, the way that these *digital signatures* work is that if Alice wants to prove her authorship of a message, she encrypts it with her *private* key.[32] Now Bob can use Alice's *public* key (which he knows because Alice has shared it with the world) to decrypt the message. Now he knows, to a high degree of certainty, that only Alice could have sent the message, because only Alice had access to the private key used to encrypt it.[33]

The final concept in this brief tour of encryption is the *cryptographic hash*,[34] an algorithm that takes input data of arbitrary size, such as a long message, and generates a much smaller *hash value*. A well-designed hash algorithm has several very nice features that collectively mean that a hash uniquely identifies the data it came from without giving away the data itself. Specifically, a good cryptographic hash satisfies the following

- Uniqueness: A given piece of data should predictably produce exactly one hash value.
- Preimage resistance: The hash value should reveal no information about the data, such that one cannot reconstruct the data from the hash.
- Collision resistance: Two different pieces of data should rarely produce the same hash value.

A hash is like a digital signature in that it provides an authenticity guarantee, except that in this case it is that only someone who had access to

32. "Authorship" here is used just to mean that the signer wishes to be identified as associated with the message. It does not mean that they necessarily authored the message in the sense that the author of a novel creates its text and has a copyright in it.

33. More advanced schemes combine standard public-key encryption and digital signatures to ensure that the message is both confidential and signed.

34. The term "one-way hash" means the same thing. A "hash function" is the mathematical process that transforms input data into a hash, and when the hash function is applied to a message, the result is sometimes called a "message digest" because the hash function has "digested" the message.

the complete text of the original message could have generated the corresponding hash. For this reason, hashes are sometimes referred to as "fingerprints"; they are small but unique identifiers.

Cryptographic hashes have a variety of uses. For one thing, they simplify the digital signature process: a message sender can encrypt just a hash of a message rather than an entire message, and others can still use the signed hash to verify the message's authorship since the hash uniquely identifies the message. Hashes can also be used to make commitments without publicly revealing the details of what one is committing to. For example, a basketball fan could publish a hash of their predictions for the March Madness NCAA bracket at the start of the tournament, and then reveal their actual bracket after the tournament is over. And finally, hashes can be used in place of content that, for whatever reason, one does not wish to store. For example, major online platforms typically scan uploaded images to see whether users are uploading known examples of CSAM. But for obvious reasons, a platform does not want to maintain a database filled with images of children being sexually exploited. By storing only the hashes of those images, the platform can still compare uploaded images to known examples of CSAM but without the legal and operational nightmares of storing actual CSAM. As we will see, hashes are particularly useful for content moderation in end-to-end encrypted systems because they allow platforms and users to make verifiable claims about content without revealing the content itself.

B. Historical Developments

Since the introduction of modern asymmetric encryption algorithms in the 1970s, law enforcement and government intelligence agencies have raised concerns that widespread private use of encryption would hamper criminal investigations and national security efforts. [35] But for many cryptographers and privacy activists, government surveillance was the principal threat that made widespread use of cryptography a moral necessity.

35. *See generally* Craig Jarvis, Crypto Wars: The Fight for Privacy in the Digital Age 111–52 (2021); Steven Levy, Crypto: How the Code Rebels Beat the Government–Saving Privacy in the Digital Age (2001). For a broader history of cryptography, see generally David Kahn, The Codebreakers: The Comprehensive History of Secret Communication from Ancient Times to the Internet (1996).

The policy debates between these two camps for and against laws restricting the use of encryption were informally dubbed the "Crypto Wars."

In the 1990s, these debates came to a head in the United States when the federal government for a time deemed strong encryption a "munition" subject to export restrictions, and lawmakers proposed "key-escrow" technologies, in which government agencies would hold special decryption keys that would enable law enforcement to decrypt communications upon receipt of a court order.[36] These skirmishes ended with defeats for the government. The export controls were relaxed to allow academic cryptographers and open-source programmer to post their work online without fear of prosecution, and the leading proposed key-escrow scheme, the Clipper Chip, collapsed in ignominy when it was shown to be insecure.

But although the 1990s left encryption legal and widely used, it was not omnipresent. In particular, most communications platforms were not end-to-end encrypted. A message from Alice to Bob would be encrypted in transit from Alice to the platform, and encrypted in storage on the platform, and encrypted in transit from the platform to Bob – but the platform held the decryption keys that would allow it to decrypt the message received from Alice, and to retrieve and decrypt the messages it stored. This meant that law enforcement could still effectively obtain unencrypted communications between platform users – either by serving the platform with legal process, or by infiltrating the platform's systems to copy out the data and decryption keys.

Following Edward Snowden's revelations of widespread national-security surveillance on Internet communications, technology companies increasingly rearchitected their communications platforms to thwart this surveillance. And if the platform itself was a potential vulnerability, the natural solution was end-to-end encryption. Now the message from Alice to Bob would be encrypted using a keypair controlled by Bob, so that the platform would have no greater ability to read the message than any random stranger would. This development led in the mid-2010s to a new wave of complaints from law enforcement about the danger of malfeasants "going dark" and the need for "back doors" for government access to en-

36. *See* Danielle Kehl et al., New Am., Doomed to Repeat History? Lessons from the Crypto Wars of the 1990s 5–12 (2015), http://newamerica.org/cybersecurity-initiative/policy-papers/doomed-to-repeat-history-lessons-from-the-crypto-wars-of-the-1990s/. *See generally* Levy, *supra* note 35.

crypted content.[37] Content moderation on end-to-end encrypted systems has arisen as a point of contention in this larger debate over strong encryption. There are at least three interlocking problems .

- First, and most seriously from law enforcement's point of view, end-to-end encryption makes it harder to detect and investigate the transmission of illegal content, such as CSAM and terrorist plots. This is a surveillance and security problem, but it is also a content-moderation problem. Identifying and removing content that has predictably harmful effects for third parties is a classic goal of content moderation.

- Second, end-to-end encryption makes it harder for platforms to defend users from spam, abuse, and haraassment. When the platform has access to all communications, it can flag specific messages that are highly likely to be unwanted and identify suspicious patterns of mass coordinated messaging. These capabilities appear to disappear when all messages are end-to-end encrypted.

- Third, content moderation for groups typically depends on delegating some of the moderation work to group admins and other users. These users make moderation decisions that are enforced by the platform. But again, when all group communications are indecipherable to the platform, it appears that it cannot effectively intervene to carry out the instructions of group administrators.

The result, as many commentators have noted, is that encrypted group communications have become vectors for abusive harassment and disinformation campaigns of the sort that platforms regularly moderate.[38]

---

37. *See* Kehl et al., *supra* note 36, at 1; Nat'l Acad. of Scis., Decrypting the Encryption Debate: A Framework for Decision Makers 7–9 (2018); Harold Abelson et al, *Keys Under Doormats: Mandating Insecurity by Requiring Government Access to All Data and Communications*, 1 J. Cybersecurity 69 (2015); Whitfield Diffie & Susan Landau, Privacy on the Line: The Politics of Wiretapping and Encryption (2d ed. 2010).

38. *See* Matamoros-Fernández, *supra* note 6, at 7–8 (citing Nic Newman et al., Reuters Institute Digital News Report 2019, at 9 (2019), https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-06/DNR_2019_FINAL_0.pdf); Cristina Tardáguila et al., *Fake News Is Poisoning Brazilian Politics. WhatsApp Can Stop It*, N.Y. Times (Oct.

Thus, the collateral harms to content moderation have become a standard argument against strong end-to-end encryption. In 2019, government officials from the United States, the United Kingdom, and Australia sent an open letter to Mark Zuckerberg, Chief Executive Officer of Facebook, calling on the company and other online platforms to "not deliberately design their systems to preclude any form of access to content"—that is, not to implement end-to-end encryption that would "severely erod[e] a company's ability to detect and respond to illegal content and activity."[39] Others have similarly argued that when encryption prevents platforms from reading users' messages, the platforms are unable to identify, respond to, and working with law enforcement on online sexual abuse of children.[40]

Privacy advocates and computer scientists, in turn, have challenged these criticisms.[41] They observe that content moderation is a suite of strategies broader than mere reading of messages: It includes user reporting workflows, automated data analysis, and message flagging mechanisms.[42] And to substantiate their claim that content moderation can be compatible with end-to-end encryption, computer science researchers have redoubled their efforts toward developing novel content moderation strategies.[43]

17, 2018), https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html.

39. *See* Letter from Priti Patel et al. to Mark Zuckerberg, Facebook, *Open Letter: Facebook's "Privacy First" Proposals* 1 (Oct. 4, 2019), https://www.justice.gov/opa/press-release/file/1207081/download.

40. *See, e.g., End-to-End Encryption*, Nat'l Ctr. for Missing & Exploited Child. (last visited Jan. 31, 2023), http://www.missingkids.org/e2ee.html.

41. *See* Mayer, *supra* note 13; Seny Kamara et al., Ctr. for Democracy & Tech., Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems (Aug. 12, 2021), https://cdt.org/insights/outside-looking-in-approaches-to-content-moderation-in-end-to-end-encrypted-systems/.

42. *See* Kamara et al., *supra* note 41, at 7–11; *see also* James Grimmelmann, *The Virtues of Moderation*, 17 Yale J.L. & Tech. 42, 55–79 (2015).

43. *See, e.g.,* Anunay Kulshrestha & Jonathan Mayer, *Identifying Harmful Media in End-to-End Encrypted Communication: Efficient Private Membership Computation*, 30 Proc. USENIX Sec. Symposium 893, 893 (2021), https://www.usenix.org/conference/usenixsecurity21/presentation/kulshrestha (noting government officials' letter to Zuckerberg as a motivation for developing privacy-preserving perceptual hash technology). To be sure, some of these content moderation technologies predate the 2019 letter. Facebook developed message franking in 2017, for example. *See* Facebook, Inc., Messenger Secret Conversations: Technical Whitepaper 11–12

II. Communication Privacy Laws

A. The Wiretap Act

Originally enacted in 1968 and as amended by the Electronic Communications Privacy Act of 1986 and other statutes,[44] § 2511 governs the interception of live communications.[45] Generally, the statute creates liable for one who "intentionally intercepts"[46] the "contents of any wire, oral, or electronic communication"[47] by means of an "electronic, mechanical, or other device,"[48] unless the interception falls under an exception in the statute.[49] The statute is typically characterized as requiring five elements:[50]

1. *Intentional*—mere ability to intercept or unintentional interception is not a violation.

2. *Interception*—the communication must be "captured or redirected,"[51] perhaps in a separate, contemporaneous transmission.[52]

(ver. 2.0 2017), https://about.fb.com/wp-content/uploads/2016/07/messenger-secret-conversations-technical-whitepaper.pdf.

44. *See* Electronic Communications Privacy Act of 1986, Pub. L. No. 99-508, secs. 101–111, 100 Stat. 1848, 1848–59.

45. *See* Wiretap Act, 18 U.S.C. § 2511. *See generally* Jim Dempsey et al., Ctr. for Democracy & Tech., An Overview of the Federal Wiretap Act, Electronic Communications Privacy Act, and State Two-Party Consent Laws of Relevance to the NebuAd System and Other Uses of Internet Traffic Content from ISPs for Behavioral Advertising 3–11 (July 8, 2008), https://cdt.org/wp-content/uploads/privacy/20080708ISPtraffic.pdf.

46. Wiretap Act § 2511(1)(a).

47. *Id.* § 2510(4).

48. *Id.* § 2510(5).

49. *See id.* § 2511(2).

50. *See, e.g., In re* Pharmatrak, Inc. Priv. Litig., 329 F.3d 9, 18 (1st Cir. 2003); Charles Doyle, Cong. Rsch. Serv., Report No. R41733, Privacy: An Overview of the Electronic Communications Privacy Act 7 (ver. 9 Oct. 9, 2012), https://crsreports.congress.gov/product/pdf/R/R41733.

51. United States v. Rodriguez, 968 F.2d 130, 136 (2d Cir. 1992).

52. *See Pharmatrak*, 329 F.3d at 21–22; United States v. Councilman, 418 F.3d 67, 80 (1st Cir. 2005) (discussing but not resolving meaning of "contemporaneous"); *cf.* Konop v. Hawaiian Airlines, Inc., 302 F.3d 868, 878–79 (9th Cir. 2002) (holding no interception to occur based on illicit viewing of website content long after content was posted).

3. *of the contents*—"information concerning the substance, purport, or meaning,"[53] and not mere metadata,[54] must be received.

4. *of an electronic communication*—including messages both as they are in transit and when in transient electronic storage.[55]

5. *using a device.*

Several exceptions to § 2511 are relevant. Law enforcement may cause the interception of otherwise protected communications, provided that certain procedural requirements are met.[56] There is no violation if one of the parties to the communication consents to the interception[57] or if the interceptor is a party to the communication.[58] Additionally, the provider of a communication service may intercept communications "in the ordinary course of its business,"[59] as a "necessary incident" to providing the service,[60] or for "the protection of the rights or property of the provider."[61]

B. The Stored Communications Act

While the Wiretap Act deals with interception of data in transit, the SCA deals with access to communications and data in electronic storage.[62] It contains two major restrictions on such stored information: a general prohibition on unauthorized access to stored communications,[63] and spe-

53. Wiretap Act § 2510(8).

54. *See, e.g.,* United States v. N.Y. Tel. Co., 434 U.S. 159, 166–67 (1977); *In re* Google Inc. Cookie Placement Consumer Priv., 806 F.3d 125, 135–39 (3d Cir. 2015).

55. *See Councilman*, 418 F.3d at 79.

56. *See* Wiretap Act § 2511(2)(a)(ii).

57. *See id.* § 2511(2)(c). Analogous state laws sometimes require all parties to the communication to consent to interception. *See* Dempsey et al., *supra* note 45, at 11 & n.37.

58. *See* Wiretap Act § 2511(2)(d).

59. *Id.* § 2510(5)(a)(ii).

60. *Id.* § 2511(2)(a)(i).

61. *Id.*

62. *See generally* Orin S. Kerr, *A User's Guide to the Stored Communications Act, and a Legislator's Guide to Amending It*, 72 Geo. Wash. L. Rev. 1208 (2004).

63. *See* 18 U.S.C. § 2701.

cific prohibitions directed to service providers.[64]

A key element of the SCA is whether a communication is in "electronic storage."[65] This term is defined narrowly, encompassing only "temporary, intermediate storage" or "storage . . . for purposes of backup protection."[66] One question is whether this definition encompasses messages that the recipient has retrieved but that still remain in the service provider's storage. A line of cases including *Theofel v. Farey-Jones* have deemed such post-transmission stored messages as "backup protection,"[67] but other courts have questioned that view.[68] Persistent cookies, by contrast, likely do not fall within this definition.[69]

The general unauthorized-access prohibition is contained in 18 U.S.C.

64. *See* § 2702.

65. *See* 18 U.S.C. § 2701(a) (flush text); *id.* § 2702(a)(1).

66. *See* Wiretap Act § 2510(17).

67. Theofel v. Farey-Jones, 359 F.3d 1066, 1077 (9th Cir. 2004); *see also* Quon v. Arch Wireless Operating Co., Inc., 529 F.3d 892, 902 (9th Cir. 2008), *rev'd on other grounds sub nom.* City of Ont. v. Quon, 130 S. Ct. 2619 (2010); O'Grady v. Superior Court, 44 Cal. Rptr. 3d 72, 84 n.9 (Ct. App. 2006); Konop v. Hawaiian Airlines, Inc., 302 F.3d 868, 879–80 (9th Cir. 2002) (assuming without deciding that posts on a website are in "electronic storage").

68. There are two lines of argument here. First, some courts posit that the SCA is limited to communications stored during transit, so post-transmission messages are outside the statute's ambit. *See* Fraser v. Nationwide Mut. Ins. Co., 135 F. Supp. 2d 623, 636 (E.D. Pa. 2001). Others observe that for services such as webmail, a post-transmission message stored on a web server is not backing up any other copy of the message. *See* Sartori v. Schrodt, 424 F. Supp. 3d 1121, 1132–33 (N.D. Fla. 2019) (quoting United States v. Weaver, 636 F. Supp. 2d 769, 772 (C.D. Ill. 2009)); *cf. Theofel*, 359 F.3d at 1076 ("An ISP that kept permanent copies of temporary messages could not fairly be described as 'backing up' those messages.").

69. *See In re* DoubleClick Inc. Priv. Litig., 154 F. Supp. 2d 497, 512–13 (S.D.N.Y. 2001); *In re* iPhone Application Litig., 844 F. Supp. 2d 1040, 1059 (N.D. Cal. 2012). *In re Intuit Privacy Litigation* denied a motion to dismiss a complaint alleging an SCA violation based on persistent cookies, but the court gave only cursory attention to whether the cookies were in "electronic storage," and the defendant did not appear to have pressed the issue. 138 F. Supp. 2d 1272, 1275–76 (C.D. Cal. 2001). *Chance v. Avenue A, Inc.* found no SCA violation for accessing persistent cookies on unrelated grounds and never reached the question of whether the cookies were in "electronic storage." *See* 165 F. Supp. 2d 1153, 1161–62 (W.D. Wash. 2001). In view of these cases, Kerr's suggestion that "several district courts have applied the SCA to regulate the placement of electronic cookies on home computers" is somewhat puzzling. Kerr, *supra* note 62, at 1214.

§ 2701. Under that statute, it is a violation to:

- *access without authorization* or exceed authorized access[70]
- *a "facility through which an electronic communication service is provided"*—generally an online server but possibly also a user device.[71]
- to obtain, alter, or prevent access to a *wire or electronic communication*[72]
- in *electronic storage*, as above.

Section 2701 further provides exceptions for access authorized by the communications service provider, by the user who sends or receives the communication, or by law enforcement.[73]

The specific provisions for service providers in 18 U.S.C. § 2702 deal with divulgence of information stored with service providers. That section creates two modes of liability depending on whether the service provider offers an "electronic communication service" or a "remote computing service."[74] Most authorities agree that a service can satisfy both definitions.[75] For example, an online email service might be an electronic communication service when it holds onto an email before the recipient reads it, but be a remote computing service after the email is read insofar as the recipient uses the email service for long-term storage.[76] If a service acts as both

70. 18 U.S.C. § 2701(a)(1)–(2).

71. *Id.* § 2701(a)(1). *Compare DoubleClick*, 154 F. Supp. 2d at 509 (suggesting that a user's personal computer on which a website cookie is stored is a "facility"), *and Chance*, 165 F. Supp. 2d at 1161 ("[I]t is possible to conclude that modern computers, which serve as a conduit for the web server's communication . . . , are facilities covered under the Act."), *with* Kerr, *supra* note 62, at 1215 & n.47 (arguing that home computers are not electronic communication services).

72. 18 U.S.C. § 2701(a) (flush text).

73. *See id.* § 2701(c).

74. *See* 18 U.S.C. § 2702(a)(1)–(2).

75. *See, e.g.*, Theofel v. Farey-Jones, 359 F.3d 1066, 1076–77 (9th Cir. 2004); Kerr, *supra* note 62, at 1215 ("[M]ost network service providers . . . can act as providers of ECS in some contexts, providers of RCS in other contexts, and as neither in some contexts as well.").

76. *See* Kerr, *supra* note 62, at 1215–16.

at the same time, then it must avoid both sets of prohibitions to escape liability.[77]

A violation of the statute requires:

- For a public electronic communication service:
  - knowingly divulging
  - the contents of a communication
  - in electronic storage,[78] or
- For a public remote computing service:
  - knowingly divulging
  - content of communication on the service
  - from (or created for) a subscriber or customer of the service
  - solely for the purpose of the storage or computing services
  - if the provider is not authorized to access the contents for other purposes.[79]

Like the Wiretap Act, § 2701 provides several exceptions to the prohibitions on services' divulging communications. No violation occurs for divulging the content of a communication:

- To the addressee or intended recipient of the communication.[80]
- With the consent of the sender or recipient.[81]

77. In *Quon*, the Ninth Circuit held that a text messaging provider was an electronic communication provider liable under § 2702(a)(1) for disclosing a police officer's messages to the city that employed the officer. *See* 529 F.3d 892, 897–98 (9th Cir. 2008). The city argued that, as the subscriber to the text service, it fell within a statutory exception that only applied to remote computing services. *See id.* at 900. In concluding that the exception did not apply, the Ninth Circuit reasoned that the text service was "more appropriately categorized as an ECS than an RCS." *Id.* at 902. Although this could be taken to mean that the two categories are mutually exclusive, a better reading is that the court held that liability could arise based on the text messaging service's electronic communication service capacity, regardless of whether it was also a remote computing service.

78. *See* 18 U.S.C. § 2701(a)(1).

79. *See id.* § 2701(a)(2).

80. *See id.* § 2701(b)(1).

81. *See id.* § 2701(b)(3).

- For forwarding the communication to its destination.[82]
- As "necessarily incident to the rendition of the service."[83]
- For "protection of the rights or property of the provider."[84]
- To law enforcement under appropriate circumstances.[85]

A different violation occurs when a public electronic communications service or remote computing service divulges non-content customer information to the government.[86] The statute offers a more limited set of exceptions to this prohibition. Non-content customer information may be disclosed with consent, as an incident of rendering the service, to protect the service provider's rights or property, or to law enforcement under appropriate circumstances.[87]

C. Pen Registers and Trap-and-Trace Devices

While the Wiretap Act deals with the acquisition of the *contents* of a communication, the PRA deals with the acquisition of *metadata* about the parties to a communication. Also known as the Pen Register Act, the statute prohibits the use of devices that record wire or electronic communications metadata without a court order.[88] For historical reasons related to telephone technology, the PRA uses two different terms to describe the regulated devices: a "pen register" records information about *outgoing* communications sent from a device,[89] while a "trap and trace device" records information about *incoming* communications sent to a device.[90] When the distinction between the two is immaterial, we will refer to them collectively as "PR/TT devices."

82. *See id.* § 2701(b)(4).
83. *See id.* § 2701(b)(5).
84. *See id.* § 2701(b)(6).
85. *See id.* § 2701(b)(2), (6-/9).
86. *See id.* § 2701(a)(3).
87. *See id.* § 2701(c).
88. 18 U.S.C. § 3121(a) ("no person may install or use a pen register or a trap and trace device without first obtaining a court order").
89. § 3127(3).
90. 18 U.S.C. § 3127(4).

For both types of devices, the key definitional phrase is that these devices capture "dialing, routing, addressing, and signaling information" (DRAS) about wire or electronic communications.[91] Despite their telephone-era names, the definitions of these tools encompass digital-era technologies, and courts have held that systems for capturing addresses in emails,[92] IP addresses,[93] and physical location information[94] can qualify as PR/TT devices.

The definitions of PR/TT devices are limited in several ways. Most importantly, a device designed to capture content is outside the scope of the Pen Register Act;[95] such a device is regulated instead by the Wiretap Act.[96] To qualify as a regulated PR/TT device, it must also collect communication metadata sent in the course of the communication, not at a different time or from a third-party data source.[97] A device used for billing purposes is exempted from the definition of pen registers, but not from the definition of trap-and-trace devices.[98] Finally, at least one court has suggested that a communication recipient's collection of metadata from the communication does not constitute operation of a trap-and-trace device.[99]

Furthermore, the statute provides several exceptions when a commu-

91. *Id.* § 3127(3) (defining a pen register as "a device or process which records or decodes [DRAS]"); *id.* § 3127(4) (defining a trap-and-trace device as "a device or process which captures the incoming electronic or other impulses which identify . . . [DRAS].").

92. *See In re* Application of the U.S. for an Ord. Authorizing the Installation & Use of a Pen Reg. & a Trap & Trace Device on E-Mail Account, 416 F. Supp. 2d 13, 16 (Dist. Court, Dist. of Columbia 2006).

93. *See* United States v. Soybel, 13 F.4th 584, 590–94 (7th Cir. 2021).

94. *See* United States v. Sanchez-Jara, 889 F.3d 418, 420 (7th Cir. 2018); United States v. Patrick, 842 F.3d 540, 543 (7th Cir. 2016).

95. *See* 18 U.S.C. § 3127(3)–(4).

96. *See, e.g., In re* Innovatio IP Ventures, LLC Pat. Litig., 886 F. Supp. 2d 888, 895 (N.D. Ill. 2012). Accidental capture of content may be permissible so long as the operator of the device "takes all reasonably available steps to minimize the collection of content information and is prohibited from making use of any content information that may be collected." *In re* Certified Question of L., 858 F.3d 591, 598 (Court of Appeals 2016).

97. *See* United States v. Fregoso, 60 F.3d 1314, 1321 (8th Cir. 1995); Brown v. Waddell, 50 F.3d 285, 291 (4th Cir. 1995).

98. *Compare* 18 U.S.C. § 3127(3), *with id.* § 3127(4).

99. *See* Captiol Recs. Inc. v. Jammie Thomas-Rasset, No. 06-cv-1497, slip op. at 8 (D. Minn. June 11, 2009) (reasoning that "the Internet could not function" if recipients could not collect metadata).

nication service provider's operation of a pen register or trap-and-trace device is automatically legal and does not require advance court authorization:

- "relating to the operation, maintenance, or testing" of the service.[100]
- relating "to the protection of the rights or property of such provider, or to the protection of users of that service from abuse of service or unlawful use of service."[101]
- "to record the fact" of a communication in order to protect the provider or a user of the service "from fraudulent, unlawful or abusive use of service."[102]
- with "the consent of the user."[103]

D. The Computer Fraud and Abuse Act

First enacted in 1984 and subsequently amended many times, the CFAA prohibits unauthorized trespass into computer systems.[104] The statute provides several pathways to a violation, the broadest of which[105] requires:

- intentional access to a *protected computer*—including any computer "used in or affecting interstate or foreign commerce or communication,"[106] which covers "every computer connected to the Internet."[107] The term "computer" is defined broadly, and an interconnected network of encrypted messaging users may constitute a "computer" under the statute.[108]

100. 18 U.S.C. § 3121(b)(1).
101. *Id.*
102. 18 U.S.C. § 3121(b)(2).
103. *Id.* § 3121(b)(3).
104. *See* 18 U.S.C. § 1030; Orin S. Kerr, *Vagueness Challenges to the Computer Fraud and Abuse Act*, 94 Minn. L. Rev. 1561, 1565–68 (2010) (summarizing history of amendments to the CFAA).
105. *See* 18 U.S.C. § 1030(a)(2)(C).
106. *Id.* § 1030(e)(2).
107. Kerr, *supra* note 104, at 1568.
108. *See* 18 U.S.C. § 1030(31); Jonathon W. Penney & Bruce Schneier, *Platforms, Encryption,*

- *without authorization or exceeding authorized access*, discussed below.
- to cause various harms, such as obtaining information from the computer,[109] obtaining value by fraud,[110] or damaging the computer in some cases.[111]

What constitutes "authorization" has been a central question for CFAA jurisprudence over the years.[112] Most recently, the Supreme Court in *Van Buren v. United States* adopted a "gates-up-or-down" approach to authorization: A person either has or lacks access to a computer or certain information on it, and contractual restrictions on how the computer or information is to be used do not affect CFAA authorization.[113] The Court entertained but did not adopt a requirement that any limit on authorization be "code-based" via technological access measures.[114] As a result, contractual and other legal notions of consent can define the scope of authorization under the statute.[115] Social norms and expectations can inform what qualifies as authorization,[116] and commentators have suggested that the use of end-to-end encryption can constitute denying authorization to third parties without the encryption keys—indeed, a code-based denial of authorization.[117]

---

and the CFAA: The Case of WhatsApp v. NSO Group, 36 Berkeley Tech. L.J. 469, 478–79 (2022); *see also* Jonathan Mayer, *The "Narrow" Interpretation of the Computer Fraud and Abuse Act: A User Guide for Applying United States v. Nosal*, 84 Geo. Wash. L. Rev. 1644, 1653–54 (2016) (noting how cloud computing requires reconceptualizing of "the scope of a computer system").

109. *See* 18 U.S.C. § 1030(a)(2)(C).

110. *See id.* § 1030(a)(4).

111. *See id.* § 1030(a)(5). As distinct from other violations of the CFAA, a violation of § 1030(a)(5) requires access "without authorization" and not exceeding of unauthorized access. *See* Int'l Airport Ctrs., LLC v. Citrin, 440 F.3d 418, 420 (7th Cir. 2006).

112. *See generally* Peter G. Berris, Cong. Rsch. Serv., Report No. R46536, Cybercrime and the Law: Computer Fraud and Abuse Act (CFAA) and the 116th Congress 6–7 (Sept. 21, 2020), https://crsreports.congress.gov/product/pdf/R/R46536.

113. Van Buren v. United States, 141 S. Ct. 1648, 1658–59 (2021).

114. *See id.* at 1659 & n.8.

115. *See* James Grimmelmann, *Consenting to Computer Use*, 84 Geo. Wash. L. Rev. 1500, 1502–03 (2016).

116. *See* Orin S. Kerr, *Norms of Computer Trespass*, 116 Colum. L. Rev. 1143, 1146 (2016).

117. *See* Penney & Schneier, *supra* note 108, at 494–95.

E. CALEA

Although the Wiretap Act permits law enforcement to intercept communications under appropriate circumstances,[118] the statute does not guarantee that the communications system will be engineered to allow for such interception. Congress recognized this problem in the early 1990s, as telephone providers transitioned from analog to digital systems incompatible with existing wiretap techniques.[119] In response to law enforcement concerns about this problem balanced against policy issues raised by privacy advocates, Congress enacted CALEA in 1994.[120]

Under the statute, telecommunications carriers must build their systems to be capable of isolating and enabling government interception of an subscriber's wire and electronic communications as well as call-identifying information.[121] The Federal Communications Commission has authority to define the scope of "telecommunications carriers" within the statutory definition of "a person or entity engaged in the transmission or switching of wire or electronic communications as a common carrier for hire."[122] However, the statute explicitly excludes from the definition "persons or entities insofar as they are engaged in providing information services."[123] An information service is defined broadly as one "offering of a capability for generating, acquiring, storing, transforming, processing, retrieving, utilizing, or making available information via telecommunications."[124] Importantly, information services include "electronic messaging services," de-

118. *See* Wiretap Act, 18 U.S.C. § 2511(2)(a)(ii).

119. *See, e.g.*, U.S. Telecom Ass'n v. Fed. Commc'ns Comm'n, 227 F.3d 450, 454 (D.C. Cir. 2000); Kristin Finklea, Cong. Rsch. Serv., Report No. R44187, Encryption and Evolving Technology: Implications for U.S. Law Enforcement Investigations 2 (ver. 5 Feb. 18, 2016), https://crsreports.congress.gov/product/pdf/R/R44187; Justin (Gus) Hurwitz, *Encryption^{Congress} mod (Apple + CALEA)*, 30 Harv. J.L. & Tech. 355, 373–76 (2017).

120. *See* Communications Assistance for Law Enforcement Act (CALEA) § 103, 47 U.S.C. § 1002; Finklea, *supra* note 119, at 2.

121. *See* CALEA § 103(a).

122. *Id.* § 102(8).

123. *See id.* § 102(8)(B)(i).

124. *Id.* § 102(6)(A). This definition is distinct from the Communications Act definition of "information service," and under current interpretations definitively encompass Internet service providers. *See* Am. Council on Educ. v. FCC, 451 F.3d 226, 232 (D.C. Cir. 2006).

fined as "software-based services that enable the sharing of data, images, sound, writing, or other information among computing devices."[125] As a result of this exception, most (but not all) Internet-based services fall outside the scope of CALEA.[126]

The FCC is also responsible for adopting technical standards for CALEA's capabilities requirements, to the extent that industry and law enforcement are unable to agree on standards independently.[127] CALEA is thus unusual among communications laws in that it gives law enforcement a hand in the technological design of systems ordinarily left to private industry.[128]

Two primary exceptions limit the applicability of CALEA.[129] First, as noted above, the statute exempts "information services."[130] Second, CALEA does not require telecommunications carriers to provide for "decrypting, or ensuring the government's ability to decrypt, any communication . . . unless the encryption was provided by the carrier and the carrier possesses the information necessary to decrypt the communication."[131] Since in an end-to-end encrypted system the "information necessary to decrypt the communication" lies solely with the communicants, this exception applies to platforms offering such encryption.[132]

125. CALEA § 102(4), (6)(B)(iii).

126. *See Am. Council on Educ.*, 451 F.3d at 234; Hurwitz, *supra* note 119, at 383–84.

127. *See* CALEA § 107(b).

128. *See* Susan Landau, *National Security on the Line*, 4 J. ON TELECOMMS. & HIGH TECH. L. 409, 412, 417–18 (2006).

129. A third exception, irrelevant to the present paper, relates to services for private networks and interconnection of carriers. *See* CALEA § 103(b)(2)(B).

130. *See id.* § 103(b)(2)(A).

131. *Id.* § 103(b)(3).

132. *See, e.g.*, *In re* Ord. Requiring Apple, Inc. to Assist in the Execution of a Search Warrant Issued by this Court, 149 F. Supp. 3d 341, 355 n.13 (E.D.N.Y. 2016); Hurwitz, *supra* note 119, at 381–82.

III. E2EE Content Moderation Proposals

A. Message Franking

Introduced as part of Facebook's secret conversations service[133] and elaborated upon in research,[134] message franking enables moderation of messages that users report to the platform as abusive or otherwise in violation of the platform's policies.[135] A challenge with moderation of user-flagged content is that, before the platform can take action against the sender of an illicit message, the platform must be able to verify who the sender was.[136] The objective of message franking is to tie the content of a message to its sender, avoiding potential false accusations and enabling the platform to take appropriate action against abuse.[137]

Digital signatures have long provided sender verifiability, but message franking aims to achieve a second objective called "deniability" or "repudiability" that digital signatures do not achieve.[138] "Off-the-record" conversations are often important for maintaining separations across social contexts: A person communicating with family and friends may not want to leave a permanent, provable record of those conversations that could wind up in the hands of coworkers, bosses, or the world.[139] A digitally signed message creates that permanent record, since anyone with access to the message sender's public key can verify the sender's authorship of the message.[140] Message franking protocols, by contrast, aim for deniability, such that no one can provably tie message contents to their senders,

133. *See* Facebook, Inc., *supra* note 43, at 11–12.

134. *See, e.g.,* Paul Grubbs et al., *Message Franking via Committing Authenticated Encryption*, 37 Proc. Ann. Int'l Cryptology Conf. 66, 67 (2017) (initiating "formal study of message franking"); Yevgeniy Dodis et al., *Fast Message Franking: From Invisible Salamanders to Encryptment*, 38 Proc. Ann. Int'l Cryptology Conf. 155 (2018).

135. For a general description of message franking, see Kamara et al., *supra* note 41, at 17–18.

136. *See* Grubbs et al., *supra* note 134, at 75.

137. *See id.*

138. *See id.*; Nikita Borisov et al., *Off-the-Record Communication, or, Why Not to Use PGP*, 2004 Proc. ACM Workshop on Priv. Elec. Soc'y 77, 79.

139. *See generally* Helen Nissenbaum, Privacy in Context: Technology, Policy, and the Integrity of Social Life (2010).

140. *See* Borisov et al., *supra* note 138, at 79.

except for the platform upon receipt of an abuse report.[141]

Message franking offers a simple but powerful mechanism for user-reported content moderation on end-to-end encrypted systems. However, the technology requires the content moderator or platform to manipulate the message's content over the wire, potentially raising legal questions.

1. Technical Overview

Message franking consists generally of steps performed by a message sender, the platform, and the recipient. To send a message, the sender first produces a cryptographic hash of the message with a randomly generated secret key.[142] This hash, the "franking tag," is sent to the messaging platform along with the message and the random key, the latter two items (but not the franking tag) being encrypted using the recipient's public key.[143] Since the random key is encrypted and thus unreadable to the platform, the franking tag at this point is meaningless content to the platform.

The platform maintains its own platform secret key, and uses that key, the franking tag, and the sender's identification to produce another cryptographic hash, the "verification tag."[144] Because the platform key is secret, third parties cannot construct false verification tags that misrepresent the senders of messages. The platform sends the encrypted content to the intended message recipient, who decrypts the content and random key. The recipient verifies that the franking tag was correctly generated; if it was not, the message is discarded as fraudulent.[145]

The verification tag, being hashed using the platform's secret key, is meaningless content to third parties, so it cannot be used to breach any sender confidentiality that the platform offers. Nevertheless, if the message recipient flags the message for moderation, then the platform can con-

141. *See* Grubbs et al., *supra* note 134, at 75.

142. *See* Facebook, Inc., *supra* note 43, at 11 (N_F and T_F are the secret key and cryptographic hash, respectively); Grubbs et al., *supra* note 134, at 75 (K_f and C_2). The variables are identified to help track the respective concepts across the different notations used in the cited references.

143. *See* Facebook, Inc., *supra* note 43, at 11; Grubbs et al., *supra* note 134, at 75.

144. *See* Facebook, Inc., *supra* note 43, at 12 (K_F and R_F are the platform key and verification tag, respectively); Grubbs et al., *supra* note 134, at 75 (K_{FB} and a).

145. *See* Facebook, Inc., *supra* note 43, at 12 ("If T_F is not verified then the recipient discards the message without displaying it.").

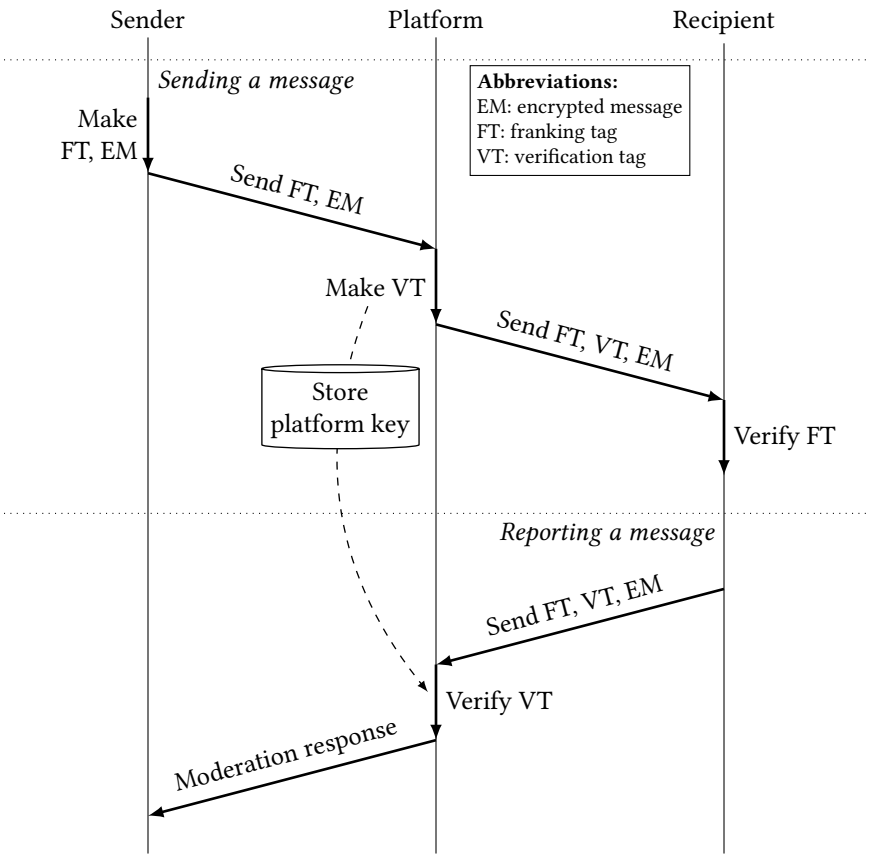

Figure 1: Diagram of communications performed in message franking.

firm the sender's identity. The recipient sends the unencrypted message, random key, sender's identification, and verification tag to the platform.[146] Along with the platform's secret key, the platform now has all the information necessary to recreate the verification tag, thereby proving that the reported message and sender identity were truthful. The platform can now take content moderation actions against the sender of the improper content, confident that the sender in fact sent that improper content.

Improvements to the above message franking system allow for greater anonymity for example,[147] or enhance performance and security of the protocols.[148] In some though not all cases, the platform may retain parts of the communication including the verification tag. First, if the recipient's device is not connected to the platform at the time, then the platform may hold onto the message and tags temporarily in the course of delivery. Second, Facebook's messaging service treats messages with attachments differently: It stores the attachment on the platform's servers to be later downloaded by the recipient, and also stores elements of the message franking communication on its servers.[149]

---

146. *See id.*; Grubbs et al., *supra* note 134, at 75 ("To report abuse, the recipient sends [the message], K_f , and a to Facebook.").

147. *See* Nirvan Tyagi et al., *Asymmetric Message Franking: Content Moderation for Metadata-Private End-to-End Encryption*, 39 Proc. Ann. Int'l Cryptology Conf. 222 (2019) (message franking on "sealed-sender" platforms where the platform is unaware of the sender's identity at the time of message delivery); Long Chen & Qiang Tang, *People Who Live in Glass Houses Should Not Throw Stones: Targeted Opening Message Franking Schemes* (Cryptology ePrint Archive, Paper 2018/994, Dec. 14, 2018), https://eprint.iacr.org/2018/994 (message franking where the recipient need not reveal the entire message contents).

148. *See* Rawane Issa et al., *Hecate: Abuse Reporting in Secure Messengers with Sealed Sender*, 31 Proc. USENIX Sec. Symposium 2335, 2339 (2022), https://www.usenix.org/system/files/sec22-issa.pdf (adding "forward" and "backward" security to message franking); Hiroki Yamamuro et al., *Forward Secure Message Franking*, 24 Int'l Conf. on Info. Sec. & Cryptology 339, 340–41 (2021) (message franking scheme resilient to compromises of platform keys); Dodis et al., *supra* note 134 (developing more efficient algorithm for message franking).

149. *See* Facebook, Inc., *supra* note 43, at 9; Dodis et al., *supra* note 134, at 161–62 (identifying a bug in Facebook's implementation of attachment storage and message franking).

2. Wiretap Act Analysis

Interception of communications implicating the Wiretap Act might occur at three points in the above message franking protocol: (1) when the platform receives the message, (2) when the recipient opens it, and (3) when the platform receives an abuse notification. Of these three points, only the first raises any substantial legal issue. On (2), the recipient opening the message does not violate the statute because the recipient is "a party to the communication."[150] And on (3), the platform receiving an abuse notification similarly has "prior consent to such interception" from the recipient,[151] and furthermore, the platform merely "listens to or copies the communication that has already been captured," which courts have held not to constitute interception under the statute.[152]

In turn, the question of whether the platform's reading and signing the franking tag of a message violates the Wiretap Act depends on four issues: (1) whether the franking tag is "content," (2) whether the platform is the intended recipient of the franking tag, (3) whether the users have consented to the franking, and (4) whether franking is in the platform's ordinary course of business.

Content. — Since the Wiretap Act is limited to interception of content, no violation would occur if the franking tag is not content. Yet that determination is surprisingly difficult. The statute defines "contents" as "any information concerning the substance, purport, or meaning" of a communication.[153] The franking tag, being a cryptographic hash of the message, carries no informational value to the platform. Yet, being derived from the content and computed so that virtually no other message would produce the same hash value, the franking tag is inextricably tied to the substance of the message and so arguably "information concerning" that message.

While scholars have recognized uncertainty in the meaning of "contents" under the Wiretap Act, the literature has largely focused on the substantive value of metadata and not whether encrypted content is con-

---

150. Wiretap Act, 18 U.S.C. § 2511(2)(d).

151. *Id.*

152. *E.g.,* Noel v. Hall, 568 F.3d 743, 749 (9th Cir. 2009). If the abuse notification is sent *before* the message recipient reads the message and the message is an audio message, then there may be an interception under the statute, as discussed *infra* Section III.D.2.

153. Wiretap Act § 2510(8).

tent.[154] The case law is no more clear. The case most on point applied the Wiretap Act, though only indirectly, to a hash of an illicit file used in a peer-to-peer filesharing network.[155] The district court found it a "closer question" whether the hash was content, and avoided it by relying on unrelated grounds for decision.[156] Another district court held that a flag identifying whether an email message was encrypted was content, "[h]owever trivial."[157] By contrast, a series of decisions about users of pirate devices for decrypting satellite television broadcasts held that no interception of content could occur by mere receipt of encrypted data unless it was decrypted.[158]

These competing views suggest that at least some courts, but not all, might be persuaded that an encrypted message qualifies as content under the Wiretap Act.[159] The encrypted message provides information about the

---

154. *See, e.g., In re* Google Inc. Cookie Placement Consumer Priv., 806 F.3d 125, 137 (3d Cir. 2015) (quoting 2 WAYNE R. LAFAVE ET AL., CRIMINAL PROCEDURE § 4.4(d) (3d ed. 2007) (content "depends entirely on the circumstances"); Orin Kerr, *Websurfing and the Wiretap Act*, WASH. POST (June 4, 2015), https://www.washingtonpost.com/news/volokh-conspiracy/wp/2015/06/04/websurfing-and-the-wiretap-act/ ("[T]he line between contents and metadata is not abstract but contextual with respect to each communication")); Steven M. Bellovin, Matt Blaze, Susan Landau & Stephanie K. Pell, *It's Too Complicated: How the Internet Upends Katz, Smith, and Electronic Surveillance Law*, 30 HARV. J.L. & TECH. 1 (2016). The latter article discusses one proposed test for distinguishing content from metadata: If data being transmitted can be encrypted without affecting transport of the data, then the data is content. *See* Bellovin, Blaze, Landau & Pell, *supra*, at 367 (discussing Shane Huang, Distinguishing Content from Metadata: The Provider-Conscious Encryption Test (May 2, 2014) (unpublished student paper)). While the authors dispute that proposed test, it is an indication that at least some scholars might accept encrypted content to be content.

155. *See* United States v. Sigouin, 494 F. Supp. 3d 1252, 1263–64 (S.D. Fla. 2019). Specifically, the defendant in the case sought to suppress evidence under the Fourth Amendment, and argued that the standard for unlawful interception under the Wiretap Act should inform the court's assessment of his reasonable expectation of privacy. *See id.* at 1263.

156. *Sigouin*, 494 F. Supp. 3d at 1264.

157. Optiver Austl. Pty. Ltd. v. Tibra Trading Pty. Ltd., No. 5:12-cv-80242, at 4 (N.D. Cal. Jan. 23, 2013).

158. *See, e.g.,* DirecTV, Inc. v. Barnes, 302 F. Supp. 2d 774, 779 (W.D. Mich. 2004).

159. For what it's worth, Congress arguably views encrypted data as content subject to the Wiretap Act, in view of a 2000 amendment to the statute requiring an annual report on the number of wiretap orders "in which encryption was encountered." Au-

sender's desire to use encryption, perhaps could be used in comparison against other messages, and certainly would reveal the sender's content if the interceptor later obtained access to the decryption keys. Although some courts might choose to follow the satellite cases in holding unintelligible encrypted messages not to be content, it would be unwise for the designer of a messaging system to assume that a court would reach that result.

Consent. — Even if the platform is not itself a party to the communication, it is possible that the sender or the recipient of the message has consented to the platform's interception of the franking tag. Such consent is a defense to Wiretap Act liability,[160] and a message platform's terms of service can suffice as consent to interception.[161] To be sure, courts have been leery of treating broad statements about data use in privacy policies or terms of service as specific consent to interception of communications, and may look unfavorably upon broad, nonspecific terms of service as evidence of users' consent to interception.[162]

However, the platform may not be able to get consent of all users. If a platform accepts messages to or from third-party services not operated by the platform, then the senders or recipients of those messages may not have provided consent to the platform's interception of franking tags or other message information.[163] One might argue that the act of transmitting the franking tag is implicit consent to platforms intercepting the tag. Because the recipient verifies the correctness of the franking tag before accepting a message for delivery, the sender cannot have a message delivered and read

---

tomatic Elimination and Sunset Reports Exemption Act, Pub. L. No. 106-97, sec. 2(a), 114 Stat. 246, 247 (2000) (codified at 18 U.S.C. § 2519, (2)(b)(iv)). The same statute added reporting on PR/TT devices but included nothing on encryption there, again suggesting that Congress viewed encrypted data as content rather than metadata. *See id.* sec. 3, 114 Stat. at 247–48.

160. *See* Wiretap Act, 18 U.S.C. § 2511(2)(a).

161. *See, e.g., In re* Yahoo Mail Litig., 7 F. Supp. 3d 1016, 1028–31 (N.D. Cal. 2014).

162. *See, e.g., In re* Google Inc. Gmail Litig., No. 13-md-2430, slip op. at 23–27 (N.D. Cal. Sept. 26, 2013); *In re* Pharmatrak, Inc. Priv. Litig., 329 F.3d 9, 20–21 (1st Cir. 2003) ("Consent 'should not casually be inferred.'") (quoting Griggs-Ryan v. Smith, 904 F.2d 112, 117–18 (1st Cir. 1990)).

163. *See Gmail*, No. 13-md-2430, slip op. at 27–28; Bruce E. Boyden, *Can a Computer Intercept Your Email*, 34 Cardozo L. Rev. 669, 678 (2012) (providing examples where consent may not be obtained).

without producing a valid franking tag.[164] The sender's request to have the message delivered, then, arguably entails authorization to generate the franking tag.

Courts have found implicit consent to interception under the Wiretap Act,[165] but have often been reluctant to do so.[166] As a result, it is not clear whether consent to interception under the Wiretap Act would be found if explicit user agreement is absent. Such situations may arise with more frequency, should the technological environment move toward interoperable messaging systems.[167]

Intended Recipient. — Even if the franking tag is content, then the platform would be permitted to intercept it in transit if the platform can show that it was "a party to the communication."[168]

Several cases illustrate the complexity of this intended-recipient exception. *In re Google Inc. Cookie Placement Consumer Privacy* involved websites that contained code instructing visitors' web browsers to transmit user information to Google, and a class of website visitors accused Google of unlawful interception.[169] The Third Circuit concluded there was no such interception because that the information was sent directly by a web re-

164. *See* FACEBOOK, INC., *supra* note 43, at 12.

165. *See In re* DoubleClick Inc. Priv. Litig., 154 F. Supp. 2d 497, 510 (S.D.N.Y. 2001) (inferring consent from "technological and commercial relationships with its affiliated Web sites").

166. *See Pharmatrak*, 329 F.3d at 20 (rejecting rule that "consent to interception can be inferred from the mere purchase of a service, regardless of circumstances"); Berry v. Funk, 146 F.3d 1003, 1011 (D.C. Cir. 1998) ("Without actual notice, consent can only be implied when the surrounding circumstances convincingly show that the party knew about and consented to the interception.") (quoting United States v. Lanoue, 71 F.3d 966, 981 (1st Cir. 1995)) (internal quotations and alterations omitted); Watkins v. LM Berry & Co., 704 F.2d 577, 581 (11th Cir. 1983) ("Consent under title III is not to be cavalierly implied.").

167. *See* Charles Duan, *A Tale of Two Interoperabilities; Or, How Google v. Oracle Could Become Social Media Legislation*, 2021 CARDOZO L. REV. DE•NOVO 246, 252–53, http://cardozolawreview.com/a-tale-of-two-interoperabilities-or-how-google-v-oracle-could-become-social-media-legislation/ (noting legislative efforts toward interoperability).

168. Wiretap Act, 18 U.S.C. § 2511(2)(d).

169. *See In re* Google Inc. Cookie Placement Consumer Priv., 806 F.3d 125, 135 (3d Cir. 2015).

quest from the visitors' browsers to Google.[170]

By contrast, in *In re iPhone Application Litigation*, mobile phones were configured to send geolocation information to Apple.[171] Even though that information was sent directly from the phone user to Apple, the district court held the § 2511(2)(d) exception inapplicable, since "[t]he intended communication is between the users' iPhone and the Wi-fi and cell phone towers," not Apple's servers.[172] In particular, Apple argued that it was the intended recipient because the phones were designed to transmit geolocation information directly to Apple.[173] The court rejected this logic on the grounds that it would allow Apple to "manufacture a statutory exception through its own accused conduct."[174]

These cases suggest divergent possible outcomes for the application of the Wiretap Act to message franking. On the one hand, a court could follow *Cookie Placement* and conclude that the franking tag is meant for the messaging platform to review and sign, making the platform the intended recipient. On the other hand, a court following *iPhone* could hold that the intended communication is between the messaging parties and not the platform. The step of sending the franking tag to the platform might be seen as manufacturing a statutory exception through protocol design. Accordingly, it is not certain that this exception would avoid Wiretap Act liability.

Business Use. — The other relevant exception is for interception "by a provider of wire or electronic communication service in the ordinary course of its business."[175] Courts have diverged greatly in their interpre-

---

170. *See id.* at 143.

171. *See In re* iPhone Application Litig., 844 F. Supp. 2d 1040, 1050–51 (N.D. Cal. 2012).

172. *Id.* at 1062; *see also In re* Pharmatrak, Inc. Priv. Litig., 329 F.3d 9, 22 (1st Cir. 2003) (holding that a violation of the Wiretap Act can be based on "[s]eparate, but simultaneous and identical, communications" with the interceptor).

173. *See iPhone*, 844 F. Supp. 2d at 1062.

174. *Id.*

175. Wiretap Act, 18 U.S.C. § 2510(5)(a). A separate, related exception relieves an employee of a communications provider from liability for interception "in the normal course of his employment while engaged in any activity which is a necessary incident to the rendition of his service or to the protection of the rights or property of the property of the provider of that service." Wiretap Act § 2511(2)(a)(i). It is unclear if this exception applies to service providers themselves. *See* Boyden, *supra* note 163, at 680.

tation of this statutory language, particularly with respect to messaging platforms' automated scanning of messages for purposes of targeted advertising.[176] Some narrowly construe "ordinary course of business" to include only interceptions that are "an instrumental part of the transmission" of a message,[177] while others more broadly apply the exception to any "customary and routine business practices" of the platform.[178]

A platform's interception of a message's franking tag is almost certainly in the ordinary course of its business under the broader construction, at least if the platform has a customary and routine business practice of content moderation as most major platforms have. The narrower construction presents a more difficult question, as message franking is not necessary to transmit messages. Nevertheless, message franking as an anti-abuse tool might be analogized to spam filtering or antivirus scanning, technologies that potentially qualify for the exception even under the narrower construction.[179]

Whether the platform stores any information from the message franking process, such as the franking or verification tags, may affect the analysis under the ordinary course of business exception. Where a platform merely has access to communications content at the time it is being transmitted and not thereafter, courts have held that the platform's access to the communication is within the ordinary course of its business.[180] By contrast, cases dealing with employer recording of telephone calls hold that such recording is not in the ordinary course of business if all calls are recorded.[181] As a result, it is possible that the baseline version of message

176. *See generally* Christopher Batiste-Boykin, *In Re Google Inc.: ECPA, Consent, and the Ordinary Course of Business in an Automated World*, 20 INTELL. PROP. L. BULLETIN 21, 30–34 (2015); Kayla McKinnon, *Nothing Personal, It's Just Business: How Google's Course of Business Operates at the Expense of Consumer Privacy*, 33 UIC J. MARSHALL J. INFO. TECH. & PRIV. L. 3, 194–200 (2018); Helen Jazzar, *Bringing an End to the Wiretap Act as Data Privacy Legislation*, 70 CASE W. RSRV. L. REV. 457, 461–69 (2019).

177. *In re* Google Inc. Gmail Litig., No. 13-md-2430, slip op. at 13 (N.D. Cal. Sept. 26, 2013); *see* Campbell v. Facebook Inc., 77 F. Supp. 3d 836, 844 (N.D. Cal. 2014) (requiring "nexus between . . . the alleged interception and the subscriber's ultimate business") (quoting *Gmail*, No. 13-md-2430, slip op. at 13).

178. *See In re* Google, Inc. Priv. Pol'y Litig., No. 12-cv-1382, at 19 (N.D. Cal. Dec. 3, 2013).

179. *See Gmail*, No. 13-md-2430, slip op. at 20 & n.4.

180. *See* Kirch v. Embarq Mgmt. Co., 702 F.3d 1245, 1250 (10th Cir. 2012).

181. *See, e.g.*, Deal v. Spears, 980 F.2d 1153, 1158 (8th Cir. 1992).

franking, which involves no retention of information, avoids a Wiretap Act violation, while more advanced versions do not.

3. SCA Analysis

There are five points in time when a communication and related franking information are stored, as relevant to the SCA: (1) on the sender's device prior to sending; (2) at the platform while the franking tag is being computed and possibly while the platform is waiting for the recipient to download the message; (3) on the platform after the recipient has downloaded the message, as a backup; (4) on the recipient's device, and (5) on the platform after the recipient has reported abuse.

Section 2701. — The general prohibition of the SCA, which covers unauthorized access to a communication service to misuse stored communications, almost certainly does not apply to any of these points in the message franking process, because all of the access to communications is likely authorized and thus not in violation of the statute.[182] If the platform has not obtained consent from the message sender for the franking process, as described above, then the platform (or the recipient) arguably lack authorization to use the sender's franking tag.[183] Even so, though, the statute permits the recipient or the platform itself to authorize access to stored content, making the sender's consent irrelevant.[184]

The sender of a message might argue, somewhat creatively, that point 1 of sending the message entails an SCA violation, to the extent that the sender did not authorize the franking protocol. The argument, akin to the *iPhone* case, would be that the sender's device is a "facility through which an electronic communication service is provided,"[185] and that the platform, through its franking-enabled messaging software, unauthorizedly accesses messages before they are sent to construct the franking tag on the sender's device. There are at least four difficulties with this argument. First, it is not clear that an individual user's device can be a "facility" for an "electronic communications service."[186] Second, at the time the franking tag is being

182. *See* 18 U.S.C. § 2701(a)(1).

183. *See supra* pp. 34–35.

184. *See* 18 U.S.C. § 2701(c)(1)–(2).

185. *Id.* § 2701(a)(1).

186. *iPhone* considered whether a user device could be considered a "facility" under the

computed, the message is arguably not in either "temporary, intermediate storage" or "backup protection," and thus fails to meet the definition of "electronic storage" as the statute requires.[187] Third and most importantly, even if the sender's device is a facility of an electronic communications service, the provider of that service (namely, the platform) can authorize access to stored communications on the device.[188] Accordingly, § 2701 is likely not violated at the time the franking tag is constructed.

Section 2702. — The second prohibition of the SCA only concerns the actions of entities providing services "to the public."[189] On the assumption that the sender and recipient do not make their devices available to the public, then only the platform's actions in points 2, 3, and 5 above are relevant to this section. Points 2 and 3 only involve disclosure of message information to the intended recipient of the message, which falls cleanly into § 2702's exceptions.[190]

To the extent that the platform reports the message and related franking information to outside authorities at point 5, the platform presumably has the message recipient's consent to report the content of the message, again falling within an exception. [191] However, the platform may be

SCA. *See* 844 F. Supp. 2d 1040, 1057 (N.D. Cal. 2012). The court held it could not for two reasons. First, such a reading would implausibly mean that "the provider of a communication service could grant access to one's home computer to third parties." *Id.* at 1058 (quoting Crowley v. CyberSource Corp., 166 F. Supp. 2d 1263, 1271 (N.D. Cal. 2001)). Second, treating the user's device as an SCA facility arguably renders the platform a "user" of that facility who can authorize the platform's access. *See id.* (discussing Chance v. Ave. A, Inc., 165 F. Supp. 2d 1153, 1161 (W.D. Wash. 2001)). A further argument against treating a user device as a facility is that "electronic communications service" is defined as one that provides services "to *users* thereof," Wiretap Act, 18 U.S.C. § 2510(15); a single-user device would not seem to fit well within that definition. *See generally* Kerr, *supra* note 62, at 1214–15 & n.47.

187. Wiretap Act § 2510(17); *see iPhone*, 844 F. Supp. 2d at 1058–59. This will depend, for example, on whether the message is placed in permanent storage on the sender's device and whether the franking tag is computed based on that permanently stored copy of the message. *See iPhone*, 844 F. Supp. 2d at 1059 ("Nor do Plaintiffs allege that Defendants accessed the data at a time when the data was only in temporary, intermediate storage.").

188. *See* 18 U.S.C. § 2701(c)(1); *iPhone*, 844 F. Supp. 2d at 1060.

189. *See* 18 U.S.C. § 2702(a)(1)–(2).

190. *See* 18 U.S.C. § 2702(b)(1).

191. *See id.* § 2702(b)(3).

barred from revealing the franking information to law enforcement. Under § 2702(a)(3), a service provider may not "knowingly divulge a record or other information pertaining to a . . . customer of such service . . . to any governmental entity." Because the franking and verification tags identify the sender of the message, those tags are information pertaining to a customer.[192] Furthermore, the message recipient's consent is irrelevant; consent must originate from "the customer or subscriber" to avoid liability under § 2702(a)(3). Unless the message sender has consented to such disclosure, the platform may only be able to reveal the franking information to law enforcement voluntarily if another statutory exception applies (to protect the service provider, to report emergencies, or to report child sexual abuse material, for example).[193]

4. PR/TT Analysis

With respect to the PRA, assuming that a message recipient's own collection of the message is exempt from the statute,[194] the only candidate for a PR/TT device is the platform's server, at the time it processes a message with a franking tag. The platform does not act as a PR/TT device upon receiving an abuse report of a communication, because any metadata the platform receives is not collected contemporaneous with the communication itself.[195]

The platform server may qualify as a trap-and-trace device, depending on the construction of "contents" described above.[196] The server captures the sender's identity in order to construct the verification tag that the platform attaches to the message, meaning that the platform server captures information "reasonably likely to identify the source" of the message.[197] However, the server also uses the franking tag, made with a hash of the

192. As discussed above, some parts of the franking information are arguably content, *see supra* pp. 32–34, and § 2702(a)(3) does not cover "contents of communications." However, there is metadata in other parts of the franking information: The verification tag includes the sender's identity information, for example.

193. *See* 18 U.S.C. § 2702(c)(3)–(5).

194. *See* Captiol Recs. Inc. v. Jammie Thomas-Rasset, No. 06-cv-1497, slip op. at 8 (D. Minn. June 11, 2009).

195. *See* United States v. Fregoso, 60 F.3d 1314, 1321 (8th Cir. 1995).

196. *See supra* pp. 32–34.

197. 18 U.S.C. § 3127(4).

message contents, to construct the verification tag. If the franking tag is considered contents of the communication, then the platform server falls outside the statutory definition.[198] Notably, a platform's encryption keys may fall within the scope of pen register interception according to one court, albeit in a case with an exceptionally unusual procedural posture.[199]

Even if the franking tag is not considered contents such that the platform server is a trap-and-trace device, the platform may nevertheless fall into one of several statutory exceptions applicable to communication service providers.[200] First, the message sender or recipient may have consented to the server's message franking activities.[201] Message franking might also fall within the exception for "operation, maintenance, and testing of a wire or electronic communication service," akin to the discussion of the business-use exception above.[202]

More importantly, the statute exempts a service provider that uses a PR/TT device for "the protection of users of that service from abuse of service or unlawful use of service."[203] Message franking, used to support content moderation, would seem to fit cleanly within this exception.

The case law supports this conclusion, though not with complete clarity. While no courts appear to have interpreted the abuse-protection provisions of the PRA, several state courts have construed state-law equivalents to the federal statute in the context of telephone caller identification services. The South Carolina Supreme Court applied that exception, holding that the caller ID service "is designed to protect the utility's subscribers from abusive or unlawful telephone calls."[204] By contrast, the Pennsylvania appellate and supreme courts did not address the abuse-protection ex-

198. *See id.*; *In re* Innovatio IP Ventures, LLC Pat. Litig., 886 F. Supp. 2d 888, 895 (N.D. Ill. 2012).

199. *See In re* Under Seal, 749 F.3d 276, 292 (4th Cir. 2014). The court primarily held that the service provider had failed to preserve the necessary arguments. *See id.* at 293.

200. *See* 18 U.S.C. § 3121(b).

201. *See supra* pp. 34–35.

202. *See* 18 U.S.C. § 3121(b)(1); *supra* pp. 36–38.

203. 18 U.S.C. § 3121(b)(1); *see also id.* § 3121(b)(2) (providing exception for use of PR/TT devices "to record the fact that a wire or electronic communication was initiated or completed in order to protect . . . a user . . . from fraudulent, unlawful or abusive use of service").

204. S. Bell Tel. & Tel. Co. v. Hamm, 306 S.C. 70, 73 (1991).

ceptions in holding caller ID services to be illegal tap-and-trace devices.[205] Although it is not clear why, the appellate decision expressed general skepticism about the service's likelihood to prevent abuse, finding it "conceivable that Caller*ID is just as likely to encourage criminal or annoying behavior as it would to discourage such conduct."[206] These decisions suggest that some courts might simply accept that content moderation is an abuse-protection objective that exempts platforms from PR/TT device regulation, while other courts might take a harder look at the platform's specific content moderation policies and practices to decide whether the exception applies.

5. CFAA Analysis

The question to be answered under the CFAA is whether the platform, in the course of the message franking protocol, accesses a protected computer in violation of the statute.[207] Since the platform has authorization to access its own servers, the sender's device and the receiver's device are the two primary computers to be considered. Per Jonathon W. Penney and Bruce Schneier, the overall encrypted messaging network could further be a protected computer.[208]

With respect to the sender's device, the argument would be that the platform, by adding message franking features to the end-to-end encrypted messaging software the sender uses, unauthorizedly accesses the sender's unencrypted message in order to generate the franking tag.[209] In a sense, the sender would argue that the message franking features are a form of spyware, working around the sender's expectation of privacy through end-to-end encryption.[210]

---

205. *See* Barasch v. Pa. Pub. Util. Comm'n, 576 A.2d 79, 297–98 (Pa. Commw. Ct. 1990), *aff'd sub nom.* Barasch v. Bell Tel. Co. of Pa., 529 Pa. 523 (1992).

206. *Id.* at 307.

207. *See* 18 U.S.C. § 1030(a)(2)(C).

208. *See* Penney & Schneier, *supra* note 108, at 488–90.

209. *See* Kerr & Schneier, *supra* note 26, at 1007–10 (describing techniques for accessing on-device plaintexts to circumvent encryption).

210. *Cf. id.* at 1009 (describing use of "government malware" to obtain IP addresses of encryption users); James Grimmelmann, *Spyware vs. Spyware: Software Conflicts and User Autonomy*, 16 Ohio State Tech. L.J. 25, 58–59 (2020) (questioning determinations of user consent when two pieces of software operate to contrary ends).

If the sender explicitly authorizes the franking tag generation (say, by accepting the platform's terms of service), then there is likely no violation of the CFAA. And even absent explicit authorization, a court might find implicit authorization based on the operation of the franking protocol.[211]

For the recipient's device, the platform installs software that, upon receipt of a message, verifies the franking tag against the sender's random key and then blocks or otherwise affects display of the message if the verification fails. If the recipient consents to the verification process, then the platform's access to the recipient's device is authorized and no violation of the CFAA occurs. Even if the recipient fails to authorize the verification process, though, no violation occurs because there is no actionable harm. The platform does not obtain any information or thing of value from the recipient's device.[212] The platform's software arguably "causes damage" to the recipient's device by blocking the "availability of data" on that device, which might violate 18 U.S.C. § 1030(a)(5).[213] But a violation of that part of the CFAA requires access "without authorization," and by virtue of installing the platform's software voluntarily, the recipient provided the necessary authorization.[214]

Yet even though the platform does not violate the CFAA for each device individually, there is a plausible argument that the platform violates the CFAA for the network as a whole. Under Penney and Schneier's network-trespass theory, the "computer" for purposes of the CFAA is the entire network of messaging participants, including the platform's servers and the devices of message senders and recipients.[215] The platform obviously is au-

211. *See supra* pp. 34–35 (describing how sender's inclusion of a franking tag might constitute implicit consent); hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1197–98 (9th Cir. 2022) (finding no lack of consent based on website's actions of making information publicly accessible); Grimmelmann, *supra* note 115, at 1508 (considering situations where a party "has done something . . . that manifests her factual consent"); *cf.* EF Cultural Travel BV v. Zefer Corp., 318 F.3d 58, 63 (1st Cir. 2003) ("[L]ack of authorization may be implicit, rather than explicit.").

212. *See* 18 U.S.C. § 1030(a)(2)–(3).

213. *See id.* § 1030(a)(5)(A), (e)(8).

214. Unlike other violations of the CFAA, damage to a computer is only actionable based on access "without authorization." *Id.* § 1030(a)(5); Int'l Airport Ctrs., LLC v. Citrin, 440 F.3d 418, 420 (7th Cir. 2006) (noting that § 1030(a)(5) does not prohibit damage from "exceeding authorized access").

215. *See* Penney & Schneier, *supra* note 108, at 490–93.

thorized to access its own network, but it is not authorized to access every piece of data thereon. After all, the whole point of end-to-end encryption is that the platform has no authorization to the communicated information.[216] Whether the platform exceeds authorized access, then, depends on what information the platform's users intend to shield the platform from viewing with their use of encryption.

There is a good argument that one piece of information that the participants intend to shield is the fact that a specific sender sent a specific piece of content. In an ordinary encrypted messaging system, the platform cannot prove that a certain user sent a particular encrypted message.[217] Even if the recipient of a message reveals the message to the platform, the platform cannot be sure who sent the message unless the sender chose to include a digital signature or some other authenticating information with the message.

With message franking, then, the platform receives new information about who sent a message, information previously inaccessible due to encryption. The spyware argument that failed with respect to the sender's device alone potentially succeeds under the network-trespass theory, because the sender can argue that the platform's software programs on the sender's and recipient's devices *together* are the "spyware" that let the platform exceed authorized access to information about the message sender's identity. Accordingly, the platform exceeds its authorized access to the encrypted messaging channel to obtain otherwise-inaccessible information, meeting all the elements of § 1030(a)(2)(C).

6. CALEA Analysis

With respect to CALEA, the question is whether a messaging platform implementing message franking would be required to implement that technology in some manner to enable interception by law enforcement, where appropriately authorized.[218] There are three candidate communications to

---

216. *See id.* at 494.

217. *See* Grubbs et al., *supra* note 134, at 67 ("But end-to-end confidentiality means that Facebook must rely on users sending examples of malicious messages. How can the provider know that the reported message was the one sent?").

218. *See* Communications Assistance for Law Enforcement Act (CALEA) § 103(a)(1), 47 U.S.C. § 1002.

which the statute might apply: (1) The transmission of an encrypted message across the platform, (2) the transmission of associated franking information with a message, and (3) a recipient's report of an abusive message sent to the platform. The first plainly falls within CALEA's encryption exception, since by definition an end-to-end encrypted platform denies the platform access to the encryption keys.[219] And assuming that law enforcement has adequate authorization to demand such interception, no technical capability is required for the third, since the platform can simply transmit to law enforcement anything it receives from the recipient's report.

However, CALEA may require covered platforms to build in interception capabilities for the message franking and verification tags as they are sent across the platform. Putting aside the threshold question of whether the platform is a "telecommunications carrier," in some situations neither of the two key exceptions of the statute apply. The encryption exception potentially does not apply because the "encryption was provided by the carrier" (the platform's software that implements franking) and "the carrier possesses the information necessary to decrypt the communication" (the platform's secret key used to encrypt the verification tag).[220] While most software-based text messaging services would likely be exempt from CALEA as "electronic messaging services,"[221] a synchronous voice-based messaging platform might be deemed sufficiently a "replacement for a substantial portion of the local telephone service" such that the platform could be deemed a "telecommunications carrier" for purposes of the statute.[222]

To the extent CALEA requires interception capabilities of platforms with message franking, what capabilities must be included? Likely the franking and verification tags would have to be retained and delivered to the government upon appropriate authorization, as that information could be deemed "call-identifying information that is reasonably available to the carrier."[223] However, by design those tags provide virtually no information without access to the unencrypted message content. Law enforcement could also obtain the sender information context that the platform uses to

219. *See* CALEA § 103(b)(3).

220. *Id.*

221. *Id.* § 102(4).

222. *Id.* § 102(8)(B)(ii); *see* Am. Council on Educ. v. FCC, 451 F.3d 226, 232–33 (D.C. Cir. 2006).

223. CALEA § 103(a)(2).

construct the verification tag; that information could overcome anonymity guarantees on platforms allowing for anonymous sending of messages.

The most concerning possibility would be that, should CALEA apply to a platform with message franking, then the government could initiate a standard-setting process at the Federal Communications Commission and propose designs for message franking that would weaken the end-to-end encryption guarantees of the messages themselves. It is unlikely that such a proposal would succeed, given that it would be an end-run around the statute's encryption exception and would also violate CALEA's warning that the statute "does not authorize any law enforcement agency or officer . . . to require any specific design of equipment, facilities, services, features, or system configurations."[224] Nevertheless, the risk of a protracted technical standards battle suggests that implementers of message franking may wish to focus on software-based messaging services and other information services clearly outside the ambit of CALEA.

B. Forward Tracing

Message franking can help platforms identify senders of reported messages, so that the platform can respond to improper content. However, platform content moderators are often concerned not just with single message transactions, but with problematic messages that are distributed widely and possibly forwarded multiple times. Where a message recipient reports the message to the platform for moderation, the platform may wish to know not just the immediate sender of that message but also all of the prior senders in the forwarding chain, to identify the origin of the message.

Forward tracing protocols overcome this limitation of message franking, helping platforms determine the originator of messages. To do this, the protocol must keep track of information about the message's forwarding path, in addition to the message content and metadata. In an end-to-end encrypted system, the challenge for such protocols is how to track this forwarding information without defeating the privacy expectations that such encryption entails.

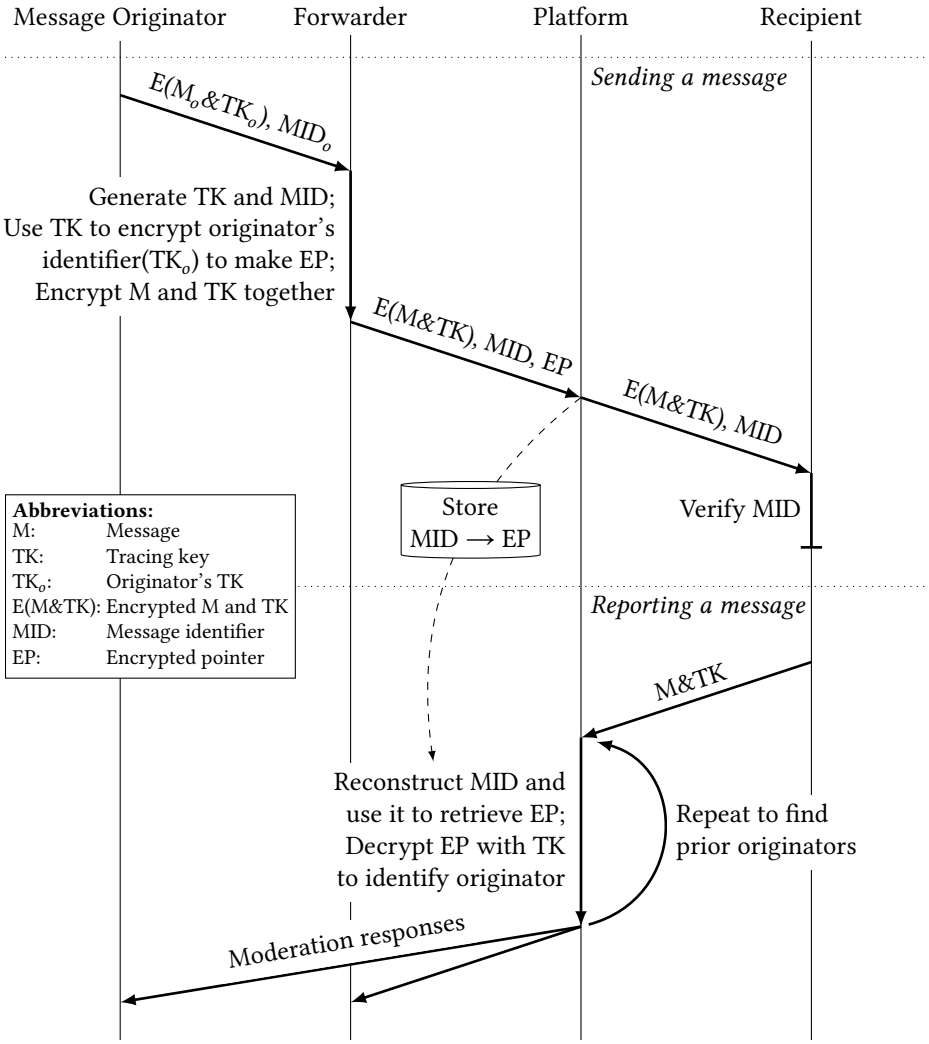

Figure 2: Diagram of communications performed in traceback.

1. Technical Overview

Nirvan Tyagi and colleagues published the basic proposal for message traceback in 2019.[225] The protocol begins much like message franking: To send a message, the sender first produces a "message identifier" based on a cryptographic hash of the message content and a randomly generated secret key, called the "tracing key."[226]

Next, the sender constructs an "encrypted pointer" that identifies the forwarding origin of the message. If the sender is forwarding another message, then the encrypted pointer is the prior message's tracing key, encrypted with the newly generated tracing key.[227] As a result, the encrypted pointer chains the forwarding path together: The tracing key of the last message in the forwarding chain can unlock the previous message's tracing key, that tracing key can unlock its predecessor, and so on.[228] If the sender has composed an original, non-forwarded message, then the sender constructs and encrypts a nonexistent tracing key, effectively cutting off the forwarding chain.[229]

The sender then encrypts the message and the tracing key together using the message recipient's public key, sending the resulting ciphertext along with the message identifier and the encrypted pointer. Upon receiving these, the platform stores the message identifier and encrypted pointer, and then delivers the message ciphertext and message identifier to the recipient. The recipient decrypts the message and verifies that the message identifier was constructed correctly before displaying the message.[230]

At this point, neither the platform nor the message recipient can determine the message forwarding chain. The message identifier is a cryptographic hash based in part on a random key, so it cannot be used to

224. *Id.* § 103(b)(1)(A).

225. *See* Nirvan Tyagi et al., *Traceback for End-to-End Encrypted Messaging*, 2019 Proc. ACM SIGSAC Conf. on Comput. & Commc'ns Sec. 413.

226. *See id.* at 416.

227. *See id.* at 417.

228. *See id.*

229. *See id.*

230. *See id.* Note that neither the platform nor the recipient verifies the encrypted pointer. Doing so would have limited value, because a determined message forwarder can always make it appear that a message has not been forwarded, for example by copying and pasting content into a new, apparently unforwarded message. *See id.* at 415.

identify related messages. The encrypted pointer contains the prior message's tracing key, meaning that it can help to identify the prior message in the chain if decrypted. But the platform lacks the tracing key to decrypt the pointer, and the recipient does not receive the pointer.[231] As a result, neither the platform nor the recipient can determine whether the message is the sender's original content or forwarded from someone else.[232]

When a message recipient wishes to report an abusive message, the recipient sends the platform the unencrypted message content and tracing key for the message. With these pieces of information, the platform can reconstruct the message identifier and thus find the associated encrypted pointer. The platform can then decrypt the encrypted pointer with the tracing key to discover the prior message's tracing key, reconstruct the prior message's identifier, find that prior message's encrypted pointer, and so on until the entire forwarding chain has been revealed.

An alternate approach, proposed by Charlotte Peale and colleagues, is to track just the original sender of a message rather than the entire forwarding chain.[233] Although it gives the platform less information to work with in responding to problematic content, this "source-tracking" approach requires no platform-side storage, and it provides message senders a greater degree of privacy then the traceback protocol.[234]

Source tracking can be understood as a variant of message franking. For all messages, the platform produces a signature analogous to the verification tag of message franking, based on the sender's identity and a message "commitment" (analogous to the franking tag).[235] The recipient of an original, unforwarded message receives and verifies the platform-produced signature.[236] When forwarding a message, however, the sender includes the platform signature inside the message to be encrypted.[237] The platform, not knowing whether the message ciphertext contains a signature

231. *See id.* at 418.

232. *See id.* at 417–18.

233. *See* Charlotte Peale et al., *Secure Complaint-Enabled Source-Tracking for Encrypted Messaging*, 2021 Proc. ACM SIGSAC Conf. on Comput. & Commc'ns Sec. 1484; Issa et al., *supra* note 148 (source tracking in sealed-sender message systems).

234. *See* Peale et al., *supra* note 233, at 1485.

235. *See id.* at 1491.

236. *See id.* at 1491 fig.3.

237. *See id.*

inside, generates a new signature, but the recipient of the forwarded message discards this generated signature and verifies only the signature found inside the message.[238] As a result, all forwards of a message will internally contain the first platform signature generated for the message, which the platform can use to identify the original sender when the message is reported.[239]

Further research improves on the basic traceback and source-tracing schemes. Tyagi and colleagues also propose a mechanism for tracing further downstream recipients who received a reported message (perhaps useful, for example, where the abusive message is a phishing scam so the platform can advise those other recipients), by having the platform store further information for downstream tracing.[240] Others extend the traceback protocol to sender-anonymous platforms.[241] James Bartusek and colleagues, End-to-End Secure Messaging with Traceability *Only* for Illegal Content (Nov. 28, 2022) (unpublished manuscript), https://eprint.iacr.org/2022/1643 proposes further mechanisms that limit the platform's ability to decrypt traceback records, and thereby discover the originator of a message, unless the message matches a database of illicit content.[242] Linsheng Liu and colleagues, *Fighting Fake News in Encrypted Messaging with the Fuzzy Anonymous Complaint Tally System (FACTS)*, in 2022 *Network & Distributed Sys. Sec. Symposium* (2022), https://www.ndss-symposium.org/wp-content/uploads/2022-109-paper.pdf similarly limits platforms' ability to uncover the original sender of a message, requiring first that a threshold number of users report the message as illicit before the sender's identity

238. *See id.* at 1491. In particular, Peale and colleagues specify giving the server a nonsense message commitment for a forwarded message, ensuring that the platform's generated signature is useless. *See id.*

239. *See* Peale et al., *supra* note 233, at 1491.

240. *See* Tyagi et al., *supra* note 225, at 420–21.

241. *See* Erin Kenney et al., *Anonymous Traceback for End-to-End Encryption*, 27 European Symposium on Rsch. Comput. Sec. 42 (2022).

242. *See* James Bartusek et al., End-to-End Secure Messaging with Traceability *Only* for Illegal Content 3–4 (Nov. 28, 2022) (unpublished manuscript), https://eprint.iacr.org/2022/1643. At a high level, the sender of a message uses the message content and a specially generated "set pre-constrained encryption" key to encrypt identity traceback information. *See id.* at 6. The corresponding private key, held by the platform, is designed based on the database of illicit content such that it can only decrypt the traceback information if the message content was contained in that database. *See id.*

can be decrypted.[243]

2. Wiretap Act Analysis

For purposes of the Wiretap Act, forward tracing protocols are largely identical to message franking protocols and so the prior analysis applies.[244] The message identifier for traceback protocols and the message commitment for source tracking are both hashes computed on the message plaintext, so they are content to the same extent that a franking tag is content.[245] The consent exception focuses on the parties to the communication at the time of transmission, so the consent of the original message sender is probably irrelevant and the exception turns on the consent of the forwarder and forward recipient.[246] The intended-recipient exception would also be analyzed in the same way that the exception was analyzed for message franking.[247]

The business use exception not apply the way it does for message franking, because the platform's purpose for forward tracing differs from that for message franking.[248] Unlike abuse prevention tools like spam filtering or antivirus scanning that courts have suggested might satisfy the exception,[249] forward tracing provides information external to an individual message transaction, namely information about who else sent or received a message with identical content. A court adopting a narrower reading of the business use exception, then, may see interception of forward tracing data as lacking a sufficient "nexus" with the platform's business purposes of message transmission.[250] This is especially so for traceback, since the plat-

243. *See* Linsheng Liu et al., *Fighting Fake News in Encrypted Messaging with the Fuzzy Anonymous Complaint Tally System (FACTS)*, *in* 2022 Network & Distributed Sys. Sec. Symposium (2022), https://www.ndss-symposium.org/wp-content/uploads/2022-109-paper.pdf.

244. *See supra* Section III.A.2.

245. *See supra* pp. 32–34.

246. *See supra* pp. 34–35; Wiretap Act, 18 U.S.C. § 2511(2)(d) (considering whether "one of the parties to the communication has given prior consent").

247. *See supra* pp. 35–36.

248. *Cf. supra* pp. 36–38.

249. *See In re* Google Inc. Gmail Litig., No. 13-md-2430, slip op. at 20 & n.4 (N.D. Cal. Sept. 26, 2013).

250. Campbell v. Facebook Inc., 77 F. Supp. 3d 836, 844 (N.D. Cal. 2014); *see Gmail*, No. 13-

form permanently stores every message's identifier regardless of whether the platform suspects wrongdoing justifying interception.[251] To the extent that message franking presented several difficult analytical questions under the Wiretap Act,[252] forward tracing enhances the difficulty of those questions.

3. SCA Analysis

As with message franking, there is likely no violation of the general unauthorized-access provision of the SCA, because all data access during a forward-tracing protocol can be authorized by the platform implementing the protocol.[253] As a result, the analysis focuses on § 2702 relating to service providers' divulgence of stored communications.

For forward tracing, information is divulged to third parties at two possible points: (1) where traceback is used and a user has reported an illicit message, the platform may provide message identifiers or encrypted pointers to law enforcement or others in the course of content moderation; and (2) where sender tracking is used, the platform's signature data is passed along the forwarding chain.

In the law enforcement reporting case (1), the disclosure of platform-stored information in the traceback protocol could meet all the requirements of § 2702(a)(1), though it is unlikely. The platform is an electronic communication service available to the public. The message and the list of users are knowingly divulged, the tracing information proving the list of users is arguably "contents of a communication," because the message identifier of the traceback protocol is cryptographically derived from the message content.[254] The more difficult question is whether the traceback data is in electronic storage. It certainly is not in temporary, intermediate storage since the platform stores it after the communication transaction has concluded.[255] And it is difficult to see how the traceback data serves

---

md-2430, slip op. at 13.

251. *See supra* pp. 37–38.

252. *See supra* Section III.A.2.

253. *See* 18 U.S.C. § 2701(a)(1); *supra* pp. 38–39.

254. *See supra* pp. 32–34.

255. *Compare* Wiretap Act, 18 U.S.C. § 2510(17)(A), *with* Tyagi et al., *supra* note 225, at 416 fig.2.

as "backup protection" even under *Theofel*, because the platform does not store that data to help users recover lost message identifiers or encrypted pointers.[256] Nevertheless, a court might be persuaded that the platform's retention of traceback data for later content moderation accountability is a kind of "backup" that meets the statutory definition.[257]

The remote computing service provision of § 2702(a)(2) probably does not apply to forward-tracing platforms because the traceback or source-tracking data is not provided to the platform solely for storage and computer processing purposes (they are intended to be forwarded to the message recipient for verification purposes).[258] However, the prohibition on disclosing customer information under § 2702(a)(3) might apply if the platform reveals traceback data to law enforcement, as that data identifies message senders so it is "information pertaining to a subscriber."[259]

Even if § 2702(a)(1) or (a)(3) applies as above, the platform may satisfy one or more of its exceptions. Message senders and recipients may have consented to the disclosure of traceback information as part of the platform's terms of service.[260] Unlike message franking, however,[261] the consent of the person reporting an improper message does not exempt disclosure of the entire forwarding chain, because other messages in the chain may not have involved the recipient at all.[262] The platform could also argue that forward tracing, as a mechanism for accountability and abuse prevention, is a necessary incident of running an encrypted messaging service

256. *See* Theofel v. Farey-Jones, 359 F.3d 1066, 1076 (9th Cir. 2004) ("But the mere fact that a copy *could* serve as a backup does not mean it is stored for that purpose."); Republic of the Gam. v. Facebook, Inc., 567 F. Supp. 3d 291, 305–06 (D.D.C. 2021) ("Facebook claims it kept the instant records as part of an autopsy of its role in the Rohingya genocide. . . . While admirable, that is storage for self-reflection, not for backup.").

257. *Cf.* Quon v. Arch Wireless Operating Co., Inc., 529 F.3d 892, 902 (9th Cir. 2008) (holding that text messaging service Arch Wireless held already-delivered text messages for "backup protection" even though it was "not clear for whom Arch Wireless 'archived' the text messages").

258. *See* 18 U.S.C. § 2702(a)(2)(B).

259. 18 U.S.C. § 2702(a)(3).

260. *See id.* § 2702(b)(3), (c)(2).

261. *See supra* p. 39.

262. For example, say that A sends a message to B who forwards it to C, and C reports the message to the platform. C's consent cannot excuse disclosing traceback information between A and B.

and protects the platform's rights and property.[263] Finally, if the disclosure is made to law enforcement to prevent serious harm or report child exploitation, then further exceptions may apply.[264]

For the sender tracking data in case (2), an argument might be that the sender's message commitment, as part of the platform's sender-tracking signature, has been divulged not just to the original message recipient but also to all third parties to whom the message is later forwarded. This argument fails both because the message recipient consents to the inclusion of the commitment and signature in future forwards, and because it is the message forwarders, not the platform itself, who are divulging those data elements.

4. CALEA Analysis

Under CALEA, a law enforcement agency could require a platform implementing a forward tracing protocol to capture message identifiers, encrypted pointers, or other information generated and transmitted to the platform in the course of the protocol.[265] As with message franking, this information is theoretically meaningless alone, but in combination with message plaintexts that law enforcement might obtain through legal or investigative means, the intercepted forward tracing information could give law enforcement access to identities of participants in a forwarded conversation chain.

For CALEA to apply, the threshold question is whether the messaging service implementing forward tracing is a telecommunications carrier under the statute, as opposed to an information service. Yet the service of forwarding messages, as required for any forward tracing protocol, would not seem to be a "replacement" for traditional telephone services, as the statute requires of a telecommunications carrier.[266] As a result, a platform would have a strong argument that the forward tracing protocol is part of an information service and thus not susceptible to the capabilities require-

263. *See* 18 U.S.C. § 2702(b)(5), (c)(3).

264. *See id.* § 2702(b)(6)–(8), (c)(4-5).

265. *See* Communications Assistance for Law Enforcement Act (CALEA) § 103(a), 47 U.S.C. § 1002.

266. CALEA § 102(8)(B)(ii); *see* Am. Council on Educ. v. FCC, 451 F.3d 226, 232 (D.C. Cir. 2006).

ments of CALEA.

If the forward-tracing platform is considered a telecommunications carrier, then the next question is whether CALEA's encryption exception exempts the platform. It probably does not. For a source-tracking protocol, the platform has the key used to encrypt the signature identifying the message source.[267] For a traceback-based protocol, the encrypted pointer is encrypted with a tracing key not in the platform's possession.[268] But on the assumption that law enforcement has obtained the tracing key from a message recipient, all that is necessary is for the platform to deliver the encrypted pointer for law enforcement to decrypt.

To the extent that the platform is deemed a telecommunications carrier, CALEA effectively places law enforcement in a privileged position above the platform itself. Like the platform, law enforcement has the ability to uncover the original sender, chain of forwarders, or the entire tree of message recipients depending on the forward tracing protocol the platform implements, so long as one plaintext message is revealed. Unlike the platform, however, law enforcement enjoys a range of compulsory legal and investigative powers to cause disclosure of that one plaintext message that unlocks the intercepted forward tracing information.

5. PR/TT and CFAA Analysis

The analysis of the PRA for forward tracing schemes is essentially identical to the analysis for message franking. The platform server is a trap-and-trace device as long as the message hashes (the message identifier for traceback, or the message commitment for source tracking) are not content, and the statutory exceptions for user consent, service operations, and abuse protection will likely apply.[269]

Analysis under the CFAA similarly mirrors that for message franking. With respect to any individual user's device, the platform has authorization to generate tracing keys or message identifiers, either based on explicit user consent or by implicit consent in order for a message to be verifiable

267. *See* Peale et al., *supra* note 233, at 1491; CALEA § 103(b)(3) (encryption exception does not apply if "the carrier possesses the information necessary to decrypt the communication").

268. *See* Tyagi et al., *supra* note 225, at 417.

269. *See supra* Section III.A.4.

upon transmission.[270] If the encrypted messaging network is considered a single "computer" for purposes of the CFAA, then the original sender of a message could argue, analogous to message franking, that the platform's software for forward tracing circumvents the privacy guarantees of end-to-end encryption and therefore exceeds authorized access.[271]

C. Server-Side Automated Content Scanning

Automated content scanning, where predefined algorithms of varying complexity sort out the good content from the bad, are a widely debated technique for content moderation. But putting aside the debate over automated moderation's effectiveness, end-to-end encryption raises a more basic question: If encryption prevents a platform from reading content, then can the platform apply algorithmic filtering in the first place? Surprisingly, it can.

1. Technical Background

This automated server-side filtering depends on a class of algorithms known as homomorphic encryption.[272] Such systems have the property that computations done on the encrypted content will predictably operate on the unencrypted content, such that the results of the computation can be retrieved once the message is decrypted.[273] The ROT-13 encryption cipher described previously is exemplifies this homomorphic property for a number of computations such as text reversal. Consider the following series of operations:[274]

$$\texttt{"GOHANGASALAMI"} \xrightarrow{\text{Encrypt}_{13}} \texttt{"TBUNATNFNYNZV"}$$
$$\texttt{"TBUNATNFNYNZV"} \xrightarrow{\text{Compute}} \texttt{"VZNYNFNTANUBT"}$$
$$\texttt{"VZNYNFNTANUBT"} \xrightarrow{\text{Decrypt}_{13}} \texttt{"IMALASAGNAHOG"}$$

270. *See supra* Section III.A.5.

271. *See supra* pp. 43–44.

272. *See, e.g.*, Nat'l Acad. of Scis., *supra* note 23, at 31; Wong, *supra* note 11, § 15.2.

273. *See* Wong, *supra* note 11, § 15.2.

274. With apologies to John Agee, who devised this palindrome.

Importantly, the fact that a platform can perform computations on homomorphically encrypted content does not mean that the platform gains any insight into the nature of the content. The content after computation appears just as scrambled and meaningless as the original encrypted content, and the result of the computation can only be perceived after decryption.[275] For this reason, homomorphic encryption enables a content scanning system to act upon problematic content without requiring message participants to expose their messages beyond the encrypted ciphertexts.

However, the actions available to the platform-based content scanner are much more limited. The scanner cannot itself determine whether content is flagged or problematic, since the results of the scanner's computations are buried within the content's encryption. Instead, the scanner can only modify the content that the recipient will see, because modifications to content are simply the results of computations on that content. For example, the scanner can attach a flag to content, or theoretically even blur or black out undesirable images. Message recipients would learn of the platform's modifications upon decrypting the messages, but the platform would not learn whose messages were flagged or modified (absent external advisement).

As a practical technology for content moderation, homomorphic encryption is far from usable on a large scale, because current algorithms are too slow to be used at the scale of large messaging platforms.[276] Nevertheless, several researchers have proposed server-side scanning systems for content moderation using homomorphic encryption.[277] Furthermore, homomorphic encryption is related to technologies known as secure multiparty communication and functional encryption, which similarly may allow a platform to perform computations on a message without giving the

275. *See* Wong, *supra* note 11, § 15.2 ("The important idea here is that the service never learns about your values and always deals with ciphertexts.").

276. *See id.* § 15.2.5 ("At the time of this writing (2021), [homomorphic encryption] operations are about one billion times slower than normal operations."); Sarah Scheffler & Jonathan Mayer, *SoK: Content Moderation for End-to-End Encryption*, 2023 Proc. on Priv. Enhancing Techs. 403, 427, https://petsymposium.org/popets/2023/popets-2023-0060.php.

277. *See, e.g.*, Song Bian et al., *Towards Practical Homomorphic Email Filtering: A Hardware-Accelerated Secure Naïve Bayesian Filter*, 24 Proc. Asia & S. Pac. Design Automation Conf. 621 (Jan. 1, 2019), https://dl.acm.org/doi/10.1145/3287624.3287699.

platform access to the message's content.[278] Future advances in cryptographic algorithms may thus create greater opportunities for platforms to moderate messages in transit that the platform cannot read.

2. Wiretap Act Analysis

For platform-based automated content scanning, the relevant point of interception occurs when the server receives the homomorphically encrypted message and performs computations on it. The message is an electronic communication under the statute, and the platform is a device that acts intentionally, so a prima facie violation of the Wiretap Act is met if the encrypted message is qualifies as "contents" and the platform "intercepts" it.[279]

Regarding interception, Bruce E. Boyden has argued that no interception should be found based on purely automated message processing, for example to append advertisements to the message.[280] However, in two Wiretap Act cases involving automated message processing (both postdating Boyden's article), the platform defendants did not raise this argument and the courts found the interception element satisfied.[281] This suggests that automated content scanning at least potentially qualifies as interception under the statute.

Turning to the definition of "contents," as discussed with respect to message franking, there is at least a plausible case that an encrypted message qualifies insofar as it is "information concerning the substance, purport, or meaning" of the message sender's communication.[282] With platform-based scanning, the argument is perhaps stronger in favor of content due to the homomorphic nature of the encryption. Even though the platform cannot discern the meaning of the message, the platform can manipulate and change the message's contents, potentially even removing

278. *See* Scheffler & Mayer, *supra* note 276, at 427–28; Théo Ryffel et al., *Partially Encrypted Deep Learning using Functional Encryption*, 32 Proc. Conf. on Neural Info. Processing Sys. 4517 (2019), https://proceedings.neurips.cc/paper/2019/hash/9d28de8ff9bb6a3fa41fddfdc28f3bc1-Abstract.html.

279. *See* Wiretap Act, 18 U.S.C. § 2511(1)(a).

280. *See* Boyden, *supra* note 163, at 702–03.

281. *See In re* Google Inc. Gmail Litig., No. 13-md-2430, slip op. at 20 (N.D. Cal. Sept. 26, 2013); *In re* Yahoo Mail Litig., 7 F. Supp. 3d 1016, 1027–28 (N.D. Cal. 2014).

282. *See* Wiretap Act § 2510(8); *supra* pp. 32–34.

information from the message. These abilities would likely make a court more inclined to treat the encrypted message as content rather than as metadata, since they suggest that the platform has a large degree of access to the sender's ability to communicate.

Assuming that the platform's receipt of the message constitutes interception of content, then the platform avoids liability under the Wiretap Act only if it meets one of the statute's exceptions: (1) if the platform is the intended recipient, (2) if users have consented to the interception, and (3) if the automated scanning is in the platform's ordinary course of business.

Regarding the first two exceptions, the analysis largely tracks that for message franking,[283] except that the argument in favor of the exceptions is potentially weaker. The common understanding of end-to-end encryption is that the platform cannot manipulate messages based on their content—something that the platform-based content scanning techniques in fact do. If users intend the platform to intercept their messages for modification or consent to the platform modifying their messages, then that intent or consent is in tension with the users' reasonable expectations of how end-to-end encryption is supposed to work. Before concluding that users have given consent, a court would likely engage in an especially searching scrutiny of a platform's terms of service before concluding that users have consented, given this tension.

By contrast, platform-side automated content scanning presents a stronger case for the Wiretap Act ordinary course of business exception. Unlike message franking, where the platform sometimes retains some of the message's content (the franking or verification tags),[284] platform-side scanning does not require the platform to retain any part of the encrypted message; indeed the platform would have little reason to do so because the modified but still homomorphically encrypted message is theoretically meaningless to the platform. Platform-side scanning thus appears akin to the spam detection or antivirus scanning practices that courts generally agree do not violate the statute.[285]

283. *See supra* pp. 34–35.

284. *See supra* pp. 37–38.

285. *See Gmail*, No. 13-md-2430, slip op. at 20 & n.4.

3. SCA Analysis

The only point in time when communication information is stored, as relevant to the SCA, is when the platform is processing the homomorphically encrypted content to augment or modify it. This is not a violation of § 2701 because the platform is the provider of the electronic communications service and so can authorize the processing.[286] Nor is it a violation of § 2702, because the result of the processed message is disclosed only to the intended recipient.[287] Server-side automated content scanning thus likely avoids liability under the SCA.

4. PR/TT Analysis

For server-side scanning to be considered a PR/TT device under the PRA, it would be necessary that the scanned, homomorphically encrypted data (1) not be content and (2) include information identifying the sender or recipient of the message.[288] This seems unlikely, given the above discussion of homomorphically encrypted data as contents under the Wiretap Act.[289] If server-side scanning does fall within the statute, then it likely satisfies the abuse-protection exceptions, given that the purpose of scanning is to flag or otherwise affect content that the platform deems abusive.[290] It likely also satisfies the service-operation and user-consent exceptions.[291]

5. CALEA Analysis

Server-side automated content scanning likely does not implicate CALEA for at least two reasons. First, the statute implicates only voice-like telecommunications services, and absent substantial technological progress it is unlikely that server-side scanning techniques will be applicable to real-time voice communications in the near future.[292] Second, it is not

286. *See* 18 U.S.C. § 2701(c)(1).

287. *See* § 2702(b)(1).

288. *See* § 3127(3)–(4).

289. *See supra* pp. 58–59.

290. *See* 18 U.S.C. § 3121(b)(1)–(2).

291. *See supra* Section III.A.4.

292. More specifically, the major current limitation of homomorphically encrypted content is that computations on such content are slow, especially when the computa-

clear what law enforcement would get out of asserting CALEA against platforms using server-side scanning. The statute only requires platforms to build in capabilities for intercepting information from communications,[293] but the intercepted information would be encrypted and CALEA does not require the platform to decrypt it. Thus, the homomorphic encryption strategies employed in server-side scanning give the platform no useful information to intercept. As a result, CALEA would probably not be asserted to require modifications to the design of a server-side automated content scanning system.

6. CFAA Analysis

As a reminder, a violation under the CFAA requires intentional unauthorized access to a protected computer that results in one of several types of harm.[294] Server-side scanning does not cause most of the types of harm enumerated in the statute. The scanning platform does not "obtain[] information" because computations on homomorphically encrypted data do not reveal information to the platform,[295] and the platform has authorization to receive the data.[296] The scanning is presumably not done "with intent to defraud" and the platform does not thereby "obtain[] anything of value"

---

tions are complex. Since automated analysis of real-time voice communications for content moderation would likely involve multilayer machine learning models, the computational cost of such models on homomorphically encrypted content would probably render this form of content moderation infeasible.

293. *See* Communications Assistance for Law Enforcement Act (CALEA) § 103(a)(1), 47 U.S.C. § 1002.

294. *See* 18 U.S.C. § 1030.

295. *See* 18 U.S.C. § 1030(a)(2)(C). One might argue that, being derived from message information, the encrypted data is information in the same way that a cryptographic hash may be "contents" under the Wiretap Act. *Cf. supra* pp. 32–34. However, the cryptographic hash has informational value to the platform in that it can be used to provably tie a message to its sender; encrypted data is ideally indistinguishable from randomness and should not serve this informational purpose. *See* Rosulek, *supra* note 23, at 22.

296. Under the "gates-up-or-down" analysis of *Van Buren*, the platform has authorization to obtain the ciphertext for purposes of delivering it; that authorization is sufficient even if the platform chooses to use the ciphertext for another unauthorized purpose such as homomorphic computation. *See* 141 S. Ct. 1648, 1658–59 (2021).

from it.[297]

The only plausible cause of action under the CFAA would be under § 1030(a)(5)(A), which prohibits the platform from "knowingly caus[ing] transmission of a program, information, code, or command, and as a result of such conduct, intentionally caus[ing] damage without authorization, to a protected computer."[298] The three candidate computers are the sender's device, the recipient's device, and the messaging system under a network-trespass theory. The program transmitted in all cases is the platform's software that performs the homomorphic encryption. That software enables the platform to perform server-side scanning, which could cause "damage" under the CFAA if the scanning results in modification of the message rather than mere addition of flagging information.[299]

Since this "damage" only occurs on the recipient's device upon receipt of a message, there is no violation with respect to the sender's device alone. With respect to the recipient's device standing alone, it would be hard to argue that the software on the *recpient's device* resulted in any "damage" because it was the software on the *client's device* that encrypted the message with a homomorphic scheme to enable server-side scanning.[300] As a result, a violation may only occur under a network-trespass theory in which the transmissions of software to *both devices* count for purposes of the statute.

Even under the network trespass theory, though, there is probably no violation of § 1030(a)(5)(A) because the damage is likely not "without authorization." Assuming that users voluntarily installed the messaging software, then the software's actions at most exceed authorized access, which this provision of the CFAA does not prohibit.[301] Furthermore, absent explicit consent, the messaging parties' choice of a platform using homomor-

297. *See* 18 U.S.C. § 1030(a)(4).

298. *Id.* § 1030(a)(5)(A).

299. *See id.* § 1030(e)(8) (defining damage as "impairment to the integrity or availability of data, a program, a system, or information").

300. This assumes that the clause in § 1030(a)(5)(A) "knowingly causes the transmission" is modified by the trailing clause "to a protected computer." It is not clear that this is the case; the text could be interpreted to mean that causing a transmission *anywhere* that results in unauthorized damage to a protected computer is a violation. If this is the correct interpretation of the statute, then the outcome is essentially the same as the outcome under the network trespass theory described here.

301. *See* Int'l Airport Ctrs., LLC v. Citrin, 440 F.3d 418, 420 (7th Cir. 2006).

phic encryption is arguably implicit authorization for the platform to use the capabilities of homomorphic encryption. Furthermore, even in non-homomorphic encryption systems, the messaging platform can block messages from being sent based on unencrypted metadata. As a result, end-to-end encryption users already should expect that platforms have some capabilities to interfere with or modify content being transmitted, mitigating the possible argument that the choice to use end-to-end encryption implies a lack of authorization for the platform to modify content.

D. Client-Side Automated Content Scanning

Even in an end-to-end encrypted system, the unencrypted plaintext message is available on the client devices that message senders and recipients use. As a result, the device itself can scan messages for illicit content without violating users' confidentiality expectations. Client-side scanning technologies have attracted significant attention recently, with Apple proposing one possible system and the U.K. Home Office funding the development of further systems.[302] Indeed, several commentators have already made initial attempts at analysis of the legality of client-side scanning.[303]

The scanning operations of client-side scanning can vary widely, from simply searching for and flagging exact matches of improper content (e.g., a wordlist-based profanity filter) to complex perceptual hashing techniques (e.g., the PhotoDNA database for detecting child sexual abuse material).[304] One might even envision using machine learning to build a client-side automatic content scanner. For our purposes, though, the legal analysis turns

302. *See* Priti Patel, *I Call on the World's Tech Giants, Please Don't Put Profit Before Safety*, The Telegraph (Sept. 8, 2021), https://www.telegraph.co.uk/politics/2021/09/08/priti-patel-call-worlds-tech-giants-please-dont-put-profit-safety/ (U.K. Home Office announcing funding for development of client-side scanning technologies for end-to-end encrypted platforms); Press Release, *Government Funds New Tech in the Fight Against Online Child Abuse* (Nov. 17, 2021) (U.K.), https://www.gov.uk/government/news/government-funds-new-tech-in-the-fight-against-online-child-abuse.

303. *See, e.g.,* Mark Rasch, *Is Apple's Client-Side Child Porn Scanning Legal?*, Sec. Boulevard (Aug. 20, 2021), https://securityboulevard.com/2021/08/is-apples-client-side-child-porn-scanning-legal/; Paul Rosenzweig, *The Law and Policy of Client-Side Scanning*, Lawfare (Aug. 20, 2020), https://www.lawfareblog.com/law-and-policy-client-side-scanning.

304. *See* Weigel, *supra* note 12; Gernand, *supra* note 12.

less on the scanning algorithms and more on the series of communications between the client device and external servers, so the communication protocols are the main focus of the description below.

1. Technical Overview

The simplest way to implement client-side scanning in an end-to-end encrypted system is to embed the entire scanner as a standalone program within the users' devices or software. Should the scanner determine that content is problematic, it can take actions that are solely limited to the client device (displaying a warning or blurring an image, for example), or it can transmit information about the determination to another computer or system via a network.[305]

Pure client-side scanning systems obviously face limitations: They are constrained by the user device's computing power, they often require periodic updates to catch new problematic content, savvy clients may be able to disable or modify the systems, and there are risks attached to giving clients access to the scanning databases and algorithms. That said, their ability to work with unencrypted content gives these systems an edge over other content moderation techniques for encrypted systems.

More advanced client-side scanning systems involve interactions with an external server. A simple option would be for the device to hash plaintext content and send the hash to a server for comparison against a database of illicit hashes.[306] Many different architectures can be imagined, but a proposed system by Anunay Kulshrestha and Jonathan Mayer demonstrates several important features and is used as an example here.

Kulshrestha and Mayer develop a client–server protocol for determin-

---

305. *See* Weigel, *supra* note 12, at 217 (noting both of these options as part of Apple's Communication Safety feature). If the client-side scanning software transmits no information beyond the device, then presumably it raises no issue under the communication privacy laws since there is no communication.

306. *See, e.g.*, Priyanka Singh & Hany Farid, *Robust Homomorphic Image Hashing*, 4 IEEE Int'l Conf. on Multimedia Info. Processing & Retrieval Workshops: Media Forensics 11 (2019), https://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Singh_Robust_Homomorphic_Image_Hashing_CVPRW_2019_paper.pdf; Matthew Green, *Can End-to-End Encrypted Systems Detect Child Sexual Abuse Imagery?*, A Few Thoughts on Cryptographic Eng'g (Dec. 8, 2019), https://blog.cryptographyengineering.com/2019/12/08/on-client-side-media-scanning/.

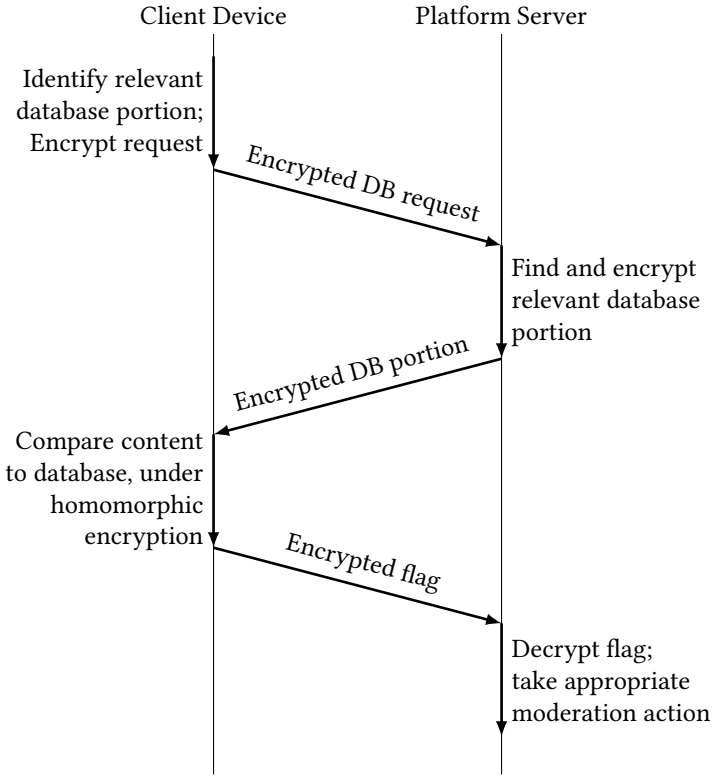

Figure 3: Diagram of communications performed in the client-side scanning system proposed by Kulshrestha and Mayer.

ing whether a piece of content is a member of a database of content (e.g., an image is within a database of contraband graphics), without revealing information about the contents of the database to users. First, the client with access to the unencrypted content requests a relevant portion of the database from the server, and the server responds by homomorphically encrypting the requested database portion and returning it to the client.[307] The client then performs a computation using the encrypted database portion and the unencrypted content.[308] This computation produces a flag indicating whether the content is the database or not.[309]

The client sends the encrypted flag back to the server, which decrypts it. Since the flag was produced by computation on the homomorphically encrypted database portion, the flag remains encrypted such that the client does not learn whether the content was flagged.[310] Furthermore, the computation is constructed such that the content itself cannot be discerned from the flag.[311] As a result, the server can take action in response to illicit content without tipping the client off, learning about clients' permissible content, or revealing information about the database of illicit content.

The above system can be thought of as a "superset" of client-side scanning, using advanced cryptographic techniques and an intricate back-and-forth communication.[312] Other client-side scanning systems use a subset of these steps. For example, many proposed systems for detecting phishing attacks or email spam involve a client homomorphically encrypting content (a URL, email, or website screenshot) and sending it to a server that computes whether the transmitted encrypted content is undesirable. The result of that computation, still encrypted, is sent back to the client where it can be decrypted and then acted upon without the server ever

307. *See* Kulshrestha & Mayer, *supra* note 43, at 899–900. The challenge is requesting the relevant portion without revealing information about the content. To do this, the client homomorphically encrypts the request for the database portion, and the server applies the request to the entire database. This results in a computed result containing only the desired database portion, but since that result is encrypted, the server does not know what portion is being returned. *See id.* at 900 (lemma 8.1).

308. *See id.* at 900–02.

309. *See id.* at 902.

310. *See id.* at 903 (theorem 11.4).

311. *See id.* (theorem 11.3).

312. *See also* Green, *supra* note 306 (discussing other complex client-side scanning systems).

learning what content was sent or even what the result of the computation was.[313] Outcome-wise, these systems seem more like server-side scanning with homomorphic encryption.[314] What distinguishes them from server-side scanning as described above, though, is that in all of them the client device accesses plaintext message content and then transmits something to a server indicating the permissibility of that content. It is this transmission of plaintext-derived information that is the key to legal analysis of client-side scanning.

2. Wiretap Act Analysis

Client-side scanning can present two activities that are potentially relevant to the Wiretap Act. The first, which we will term a "database request," involves the client device requesting from the platform information about the scanning algorithm, such as a relevant fragment of a matchlist database. In the second, termed "flag computation" here, the client device performs scanning computations on received content and possibly transmits the computation results back to the platform. Various scanning systems may involve only one or none of these activities, and they may implement the activities either on message senders' devices or on recipients' devices.

For both activities above, the threshold question under the Wiretap Act is whether either constitutes an interception. Generally, only acts contemporaneous or at least close in time with the overall sending of communications qualify under the statute.[315] So, for example, if the device waits to perform the flag computation after the message has been sent, then the flag

313. *See* Edward J. Chou et al., *Privacy-Preserving Phishing Web Page Classification via Fully Homomorphic Encryption*, 2020 IEEE Int'l Conf. on Acoustics Speech & Signal Processing 2792, 2793 & fig.1, https://ieeexplore.ieee.org/abstract/document/9053729 (applying homomorphic encryption to screenshots of phishing attack websites); Imtiyazuddin Shaik et al., *Privacy Preserving Machine Learning for Malicious URL Detection*, 2021 Database & Expert Sys. Applications 31, 32 (same for URLs); Trinabh Gupta et al., *Pretzel: Email Encryption and Provider-Supplied Functions Are Compatible*, 2017 Proc. Conf. ACM Special Int. Grp. on Data Commc'n 169, 169–70 (same for email spam detection).

314. *See supra* Section III.C.

315. *See* Noel v. Hall, 568 F.3d 743, 749 (9th Cir. 2009); Konop v. Hawaiian Airlines, Inc., 302 F.3d 868, 878 (9th Cir. 2002).

computation would not count as "interception."[316] This timing distinction is notable because some client-side scanning systems, such as one Apple proposed, require scanning to occur at the time content is received.[317]

Assuming that the database request and/or flag computation steps qualify as interceptions, they satisfy a prima facie case under the Wiretap Act if the information intercepted is content.[318] The flag computation undoubtedly involves content, because the result of the computation indicates whether the user's message is illicit under the automated scanner's content moderation standards.[319] The database request could also be content if the request depends on information in the message, as in Kulshrestha and Mayer's proposed system.[320] To be sure, in that system the request is homomorphically encrypted such that the platform cannot discern anything about the message from it. So whether the database request is content will turn on the question of whether homomorphically encrypted content is content, as discussed above.[321]

To the extent that a potential Wiretap Act violation is present, we next turn to the statutory exceptions. The intended-recipient exception likely does not apply where the server or a third-party system learns the result of the client-side scan.[322] However, there is a strong case for the business use

316. The device of course must retain the message in order to perform the flag computation on it later, and that retention would qualify as interception under the Wiretap Act. But so long as the user of the device chooses to retain a copy of the message, that retention is almost certainly consented to.

317. *See* Abhishek Bhowmick et al., The Apple PSI System 5 (July 29, 2021) (unpublished manuscript), https://www.apple.com/child-safety/pdf/Apple_PSI_System_Security_Protocol_and_Analysis.pdf ("All client-side processing of [incoming message data] must be done as soon as the client receives this [data]."). There is a good policy justification for making the timing of client-side scanning relevant to Wiretap Act liability. Where the flag computation occurs well after a message is sent or received, the user has the option of deleting the message to avoid the scanner, whereas the user has no choice when the computation is contemporaneous with the message transmission. *Cf. id.* ("[A] device reset during the delay period would cause the [message data] to never be processed.").

318. *See* Wiretap Act, 18 U.S.C. (1)(a).

319. *See* Kulshrestha & Mayer, *supra* note 43, at 903.

320. *See id.* at 900.

321. *See supra* pp. 58–59.

322. *See* Wiretap Act (2)(d).

exception.[323] As discussed with respect to the extension phone cases, the business use exception can turn on whether every communication is being intercepted, as opposed to only communications involving illicit activity.[324] Some client-side scanning systems such as Kulshrestha and Mayer provide the platform with information only about flagged content, thus falling potentially within the scope of the exception as long as the grounds for flagging are tied to a legitimate business objective.

The consent exception presents an even greater tension between explicit terms of service that might authorize client-side scanning on the one hand, and expectations of end-to-end encrypted systems on the other.[325] Users might argue that by choosing to use end-to-end encryption, they communicated an intention user to prevent others—specifically including the platform—from learning about their content.[326] Furthermore, an end-to-end encrypted platform likely holds itself out as being unable to learn substantive information about users' messages.[327] A court determining whether users have consented to client-side scanning would thus arguably have to reconcile a conflict between explicit contractual terms and context of the service and its users' expectations.[328]

---

323. *See id.* (2)(a).

324. *See supra* pp. 37–38.

325. *Cf. supra* p. 59.

326. *See* Rasch, *supra* note 303 (arguing that Apple's terms of service are insufficient to authorize client-side scanning); Erica Portnoy, *Why Adding Client-Side Scanning Breaks End-to-End Encryption*, Elec. Frontier Found. (Nov. 1, 2019), https://www.eff.org/deeplinks/2019/11/why-adding-client-side-scanning-breaks-end-end-encryption ("Client-side scanning mechanisms will break the fundamental promise that encrypted messengers make to their users: the promise that no one but you and your intended recipients can read your messages or otherwise analyze their contents to infer what you are talking about."); *see also* Jeffrey Vagle, *Client-Side Scanning: A New Front in the War on User Control of Technology*, Just Sec. (Oct. 28, 2021), https://www.justsecurity.org/78749/client-side-scanning-a-new-front-in-the-war-on-user-control-of-technology/ (noting implicit expectation of control over data based on ownership of the device on which it is stored).

327. *See, e.g., Signal* (last visited May 23, 2023), https://www.signal.org/ ("We can't read your messages or listen to your calls, and no one else can either.").

328. *See* Ronald J. Gilson et al., *Text and Context: Contract Interpretation as Contract Design*, 100 Cornell L. Rev. 23, 49–53 (2014) (describing textualist and contextualist schools of contract interpretation).

3. SCA Analysis

The relevant points of data processing for the SCA are when (1) the user's device requests database information about illicit content, (2) the user's device scans message content, and (3) the user's device sends a report of flagged content to the server.

Section 2701. — None of these points in time likely gives rise to a violation of this statutory prohibition. The platform itself, as the provider of the electronic communications service over which messages are scanned, can authorize access to those communications.[329] Similar to message franking, the user might argue that the user's device itself is a "facility through which an electronic communication service is provided" under § 2701(a). This argument likely fails for the reasons given above with respect to message franking.[330]

Section 2702. — For purposes of this analysis, it is assumed that the client-side scanning takes place within platform-provided software on the user's device. Points 1 and 3 could trigger liability under this section to the extent that the illicit content database request and/or the flagged content report go to a third party external to the platform. Assuming that the database request contains some (possibly encrypted) content relating to the information to be scanned client-side,[331] then the question to be analyzed is largely the same for both points: whether divulging a part of the user's content, in the process of a client-side scanning protocol, violates § 2702.

If what is being scanned is a message or communication, then the service is an electronic communication service.[332] Assuming that the service is public, then a violation of § 2702(a)(1) occurs if the message is in "electronic storage." If the message is scanned before it is read, then the message is in "temporary, intermediate storage" that meets the statutory definition.[333] If the message is scanned after it is read, then under Ninth Circuit

---

329. *See* 18 U.S.C. § 2701(c)(1).

330. *See supra* pp. 38–39 and notes.

331. If the request is encrypted, as in the Kulshrestha and Mayer protocol, then the database request is content to the extent that encrypted information is content as described above. *See supra* pp. 32–34. If it is not content, then there could be a violation of § 2702(a)(3) if the database is run by a government entity.

332. *See* Wiretap Act, 18 U.S.C. § 2510(15).

333. *See* Wiretap Act § 2510(17)(A).

precedent at least, the message could be considered in "backup protection" storage that also meets the definition.[334] However, since the unencrypted and scanned message on the user's device is probably the user's only copy of the message, there is a good argument that the message is not a backup and so is not in electronic storage.[335]

The platform could also be a remote computing service, for example if the user device is using the platform for cloud storage.[336] In this case, liability under § 2702(a)(2) arises if the platform "is not authorized to access the contents" for purposes other than backup and computing. This will depend on whether the platform's terms of service provide sufficient authorization to divulge information. As with the Wiretap Act discussion above, the platform's offering of end-to-end encryption is arguably in tension with the notion that the user has authorized the platform to disclose content to third parties, perhaps making a court more inclined to find a violation here.

Exceptions to § 2702(a) potentially do not apply. No liability occurs, for example, where the message sender or recipient (or the remote computing service subscriber) consents to content being divulged as part of a client-side scanning protocol, but as noted above, such consent is at odds with expectations of end-to-end encryption. The platform could also argue that client-side scanning is a necessary incident of the platform's service or protects the platform's rights and property.[337] Some courts have treated this exception as giving platforms broad discretion,[338] but by analogy to the business use exception under the Wiretap Act, other courts may be inclined to read this exception narrowly to incidents of message transmission or storage, depending on the nature of the automated scanning.[339] The platform may also be able to use exceptions for disclosure to law enforcement to prevent specific harms,[340] which could justify divulging the final determination of illicit content but not divulging information at the

334. *See id.* § 2510(17)(B); Theofel v. Farey-Jones, 359 F.3d 1066, 1077 (9th Cir. 2004).

335. *See* Sartori v. Schrodt, 424 F. Supp. 3d 1121, 1133 (N.D. Fla. 2019); Republic of the Gam. v. Facebook, Inc., 567 F. Supp. 3d 291, 305 (D.D.C. 2021).

336. *See* 18 U.S.C. § 2711(2).

337. *See* § 2702(b)(5).

338. *See Facebook*, 567 F. Supp. 3d at 309.

339. *See supra* pp. 36–38.

340. *See* 18 U.S.C. § 2702(b)(6)–(8).

database request stage (point 1 above).

4. PR/TT Analysis

Either the client device software or the platform server could qualify as a PR/TT device under the PRA. The determinative question is what information is transmitted during the scanning process—database requests, content hashes, or encrypted flags, for example. That information may be deemed content, exempting the client device or platform server from the statute.[341] Even if it is not content, that information may not identify the sender or recipient of any communication, again leaving the client device or platform server outside the ambit of the statute.[342]

To the extent that a client-side scanning system does qualify as a PR/TT device, the user-consent, platform-operation, or abuse-protection exceptions of the PRA may apply.[343] Application of those exceptions will largely track the message franking analysis,[344] with a few notes. As noted above, user consent to scanning is potentially in tension with expectations about end-to-end encryption, so that exception to the PRA may apply with less force.[345] And the abuse-protection exceptions may depend on whose device the automated scan occurs on.[346] Scanning content before displaying it to the recipient of a message probably presents a strong case for protecting the recipient from abuse of the service. Where a user's own content is being scanned, however, it is less plausible that client-side scanning is protecting those users from their own abuses; the platform would have to argue that client-side scanning protects the user community at large.[347]

---

341. *See* 18 U.S.C. § 3127(3)–(4); *supra* pp. 32–34.

342. *See* 18 U.S.C. § 3127(3)–(4).

343. *See* 18 U.S.C. § 3121(b).

344. *See supra* Section III.A.4.

345. *See* 18 U.S.C. § 3121(b)(3).

346. *See id.* § 3121(b)(1)–(2).

347. The statute seems open to this interpretation: The exception provides for using a PR/TT device for the "protection of users of that service" generally. *Id.* § 3121(b)(1).

5. CALEA Analysis

A client-side automated content scanning system could be usefully modified to aid law enforcement in intercepting communications. If the system is already configured to report flagged content back to the platform, then law enforcement might demand that the platform update its matching databases or algorithms to flag certain communications of interest to law enforcement. For example, if law enforcement suspected a certain message platform user of money laundering, then it could ask the platform to flag the phrase "money laundering" in that user's messages as illicit content within the client-side scanning system, such that the platform would be notified of such messages and could forward that information on to law enforcement. Kulshrestha and Mayer have identified this possibility of law enforcement manipulation of client-side scanning algorithms, and are developing technical strategies to act as "canaries in the coal mine," revealing whether the moderation policies have been modified.[348]

The question to be addressed here is whether law enforcement could compel such modifications under CALEA. The threshold limitation is that the statute only applies to voice-like telecommunications services.[349] Automated scanning programs are again unlikely to be implemented as a technical matter for real-time voice communications for a number of reasons relating to computational speed. First, these scanning programs would be executed on users' phones or home computers, which have limited processing power.[350] Second, to the extent that the scanning program involves homomorphic encryption, the computational complexity likely renders real-time scanning infeasible as discussed above.[351]

If technology for real-time scanning of voice communications does become available, though, then there is a good argument that law enforcement could require platforms adopting that technology to build in capabil-

348. *See* Kulshrestha & Mayer, *supra* note 43, at 905 ("The server could also collaborate with trusted third parties (e.g., civil society groups) to validate the hash set . . . ."); Sarah Scheffler et al., *Public Verification for Private Hash Matching*, 2023 Proc. IEEE Symposium on Sec. & Priv. 2074.

349. *See* Communications Assistance for Law Enforcement Act (CALEA) § 102(8)(B)(ii), 47 U.S.C. § 1002.

350. *See* Rosenzweig, *supra* note 303 (noting that client-side scanner "would cause increased processor usage and, thus, decreased battery life").

351. *See generally* Green, *supra* note 306.

ities for flagging content of interest to law enforcement. Such capabilities would not violate CALEA's encryption exception, as the computation and transmission of flagging information occurs after the content has been decrypted. To the extent that a client-side automated content scanning system works for voice communications and returns information to the platform, CALEA could be used to require the platform to modify the scanning system's content moderation policies.

The "canary" technologies that reveal whether the government has modified the content moderation policies[352] present a further problem. Under CALEA, regulated communications services must "protect[] . . . information regarding the government's interception of communications and access to call-identifying information."[353] So assuming that CALEA imposes technical capability requirements on client-side scanning systems, the statute might further prohibit technologies that reveal modifications to the scanning policies.

6. CFAA Analysis

Because client-side scanning provides a platform or other entity with information from a user's computer, it potentially violates the CFAA's prohibition on unauthorizedly obtaining information from a protected computer.[354] The user's device is the protected computer under the statute,[355] so the dispositive issues are (1) whether transmitting a hash or flag back to the platform constitutes obtaining information under the statute, and (2) whether the platform had authorization to access the user's computer via client-side scanning software.

Regarding the first issue, the question of whether a hash or flag constitutes "information" under the CFAA parallels the questions regarding "contents" under the Wiretap Act.[356] Unlike the latter statute, however, the CFAA offers no definition of "information."[357] The few cases to have

352. *See supra* note 348.

353. *See* CALEA § 103(a)(4)(B).

354. *See* 18 U.S.C. § 1030(a)(2)(C).

355. *See* 18 U.S.C. § 1030(e)(2)(B).

356. *See supra* pp. 32–34.

357. *See* Orin S. Kerr, *Focusing the CFAA in Van Buren*, 2021 Sup. Ct. Rev. 155, 160 (characterizing interpretation of "obtains . . . information" as an open question under the

considered the statutory phrase have focused on what it means to "obtain" information, not the nature of the information itself; nevertheless these cases at least suggest that courts are likely to interpret "information" broadly.[358] Since hash values and flags do indicate something about the content on the user's computer, they could comfortably fit within such a broad interpretation of "information."[359] The strongest argument to the contrary would probably be to analogize to the Fourth Amendment, where several scholars have vigorously argued that hash values merely indicating the presence of contraband are not searches.[360] Yet putting aside the question of whether the CFAA reaches further than the Fourth Amendment, other scholars contend that hash matching is indeed a Fourth Amendment search, bolstering the view that hashes are information under the CFAA.[361]

Regarding unauthorized access, the platform would point to the user's voluntary installation of software or purchase of the device with client-side scanning software on it. It could also rely on its terms of service to show that the user authorized client-side scanning. On the other hand, as with consent under the Wiretap Act, the platform's promoting itself as an end-to-end encrypted system and users' expectations of such encryption might constitute a refusal of authorization for such scanning.[362]

CFAA).

358. *See, e.g.,* Am. Online, Inc. v. Nat'l Health Care Disc., Inc., 121 F. Supp. 2d 1255, 1276 (N.D. 2000) ("mere observation of the data" is sufficient for violation of § 1030(a)(2)(C)) (quoting legislative history); United States v. Drew, 259 F.R.D. 449, 457 (C.D. Cal. 2009) (noting that the intentionality and obtaining-information elements of § 1030(a)(2)(C) "will always be met when an individual using a computer contacts or communicates with an Internet website").

359. While the numerical content of a hash ideally conveys no information about the underlying content, the fact that two hash values match each other strongly indicates that the underlying content is the same, which is "information" in the sense that it reduces uncertainty about the state of the world.

360. *See, e.g.,* Richard P. Salgado, *Fourth Amendment Search and the Power of the Hash*, 119 HARV. L. REV. F. 38, 42 (2005); Wei Chen Lin, Note, *Where Are Your Papers?: The Fourth Amendment, the Stored Communications Act, the Third Party Doctrine, the Cloud and Encryption*, 65 DEPAUL L. REV. 1093, 1118–19 (2016).

361. *See, e.g.,* Dennis Martin, Note, *Demystifying Hash Searches*, 70 STAN. L. REV. 691, 626–27 (2018); Denae Kassotis, *The Fourth Amendment and Technological Exceptionalism After Carpenter: A Case Study on Hash-Value Matching*, 29 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 1243, 1313–14 (2019).

362. *See* Rosenzweig, *supra* note 303 ("Aspects of the Computer Fraud and Abuse

As one of us has argued, such questions about the scope of user consent under the CFAA are both factual and normative.[363] From a factual perspective, the question would be whether the platform's terms of service and other disclosures are sufficiently clear to make users actually understand that the client-side scanning software undercuts the end-to-end encryption as to the platform. But a court taking a more normative point of view would ask whether users ought to tolerate hashing or flagging of their content via client-side scanning, and incorporate that normative expectation into the meaning of authorization under the CFAA.[364]

IV. Discussion

A. Statutory Ambiguities and Proposed Amendments

The foregoing statutory analysis found that the content moderation technologies in question would likely survive under the communication privacy laws considered. Yet our conclusions are not drawn with strong certainty because of numerous statutory ambiguities encountered in our legal analysis. That communication privacy law is "famous (if not infamous) for its lack of clarity" is not new.[365] But the ambiguities that these new technologies face go beyond run-of-the-mill interpretive questions, reflecting tensions between novel cryptographic techniques and the decades-old communications paradigms that the statutes assume.

In particular, our analysis has identified several "statutory components"—elements of the communication privacy laws that are largely but not quite the same, and that consistently pose challenges in view of the content moderation technologies studied. Almost every statute includes a consent component, for example, but the precise rules

Act . . . might be read to prohibit [client-side scanning].").

363. *See* Grimmelmann, *supra* note 115.

364. *See id.*; *cf.* Kerr, *supra* note 116 (arguing that courts should look to commonly shared norms among computer users in making such determinations).

365. Steve Jackson Games, Inc. v. U.S. Secret Serv., 36 F.3d 457, 462 (5th Cir. 1994) (citing Forsyth v. Barr, 19 F.3d 1527, 1542–43 (5th Cir. 1994)); *see* Ariana R. Levinson, *Toward a Cohesive Interpretation of the Electronic Communications Privacy Act for the Electronic Monitoring of Employees*, 114 W. Va. L. Rev. 461, 463–64 (2011) ("The ECPA has been described by experts as dense, intricate, and difficult for lawmakers, lawyers, and even scholars to interpret.") (citing sources).

differ across the Wiretap Act, the SCA, the PRA, and the CFAA (where it is instead called "authorization"). Our review of these components below addresses two questions: first, how they could better be interpreted in light of technological change; and second, why they are inconsistent across statutes and whether a unified, modular definition across the laws would be preferable.

Left unaddressed, these ambiguities could have the unfortunate ironic consequence that the communication privacy laws unintentionally work to reduce privacy. Online platforms have legal, ethical, and business incentives to moderate content.[366] A messaging platform that hopes to moderate users' messages, then, faces a choice: either adopt end-to-end encryption and implement new technologies for content moderation, or eschew encryption and moderate content by traditional means. To the extent that those new technologies are legally risky due to interpretive ambiguities, the platform may find the latter path safer, and that platform's users would not enjoy the privacy benefits of encrypted messaging.

1. Information and Content

The treatment and taxonomization of information plays a role in all of the communication privacy laws. The Wiretap Act, the SCA, and the PRA draw a distinction between the "contents" of communications and metadata indicating call routing or origination.[367] CALEA draws distinctions between "wire and electronic communications" on the one hand, and "call-identifying information" on the other.[368] The CFAA similarly proscribes "obtain[ing] information" beyond authorization, although "information" is left undefined.[369]

The content/metadata distinction generally does not play well with the data structures produced by modern cryptographic algorithms, because it

366. *See, e.g.*, James Grimmelmann & Pengfei Zhang, *An Economic Model of Intermediary Liability*, 37 Berkeley Tech. L.J. (forthcoming 2023) (manuscript at 17), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4422819; Klonick, *supra* note 1, at 1625–30; Felix T. Wu, *Collateral Censorship and the Limits of Intermediary Immunity*, 87 Notre Dame L. Rev. 293, 301 (2011).

367. *See* Wiretap Act, 18 U.S.C. § 2510(4); § 2702(a)(1)–(2); § 3127(3)–(4).

368. *See* Communications Assistance for Law Enforcement Act (CALEA) § 103(1)–(2), 47 U.S.C. § 1002.

369. *See* 18 U.S.C. § 1030(a)(2)(C).

is uncertain whether "contents" under the Wiretap Act encompasses data cryptographically derived from a message, such as a hashed franking tag or homomorphically encrypted content.[370] Similarly, with respect to the CFAA, there is at least a plausible question as to whether a platform "obtains . . . information" from a user's computer when the platform receives a hash or flag indicating the illicitness of client-side scanned content.[371]

These uncertainties, which commentators have noted in other legal and technological contexts,[372] reflect a generational divide between the statute and the technology. In 1986 when the Electronic Communications Privacy Act amended the Wiretap Act to address electronic communications,[373] it would have been reasonable to assume that any content that could be usefully intercepted was readable plaintext. Encryption was known at that time,[374] but the expectation was that encrypted messages were "unintelligible" without the decryption keys.[375] There would have been little value to addressing the legal ramifications of intercepting encrypted content that had no informational use without the encryption keys.

Cryptographic hashes and homomorphic encryption, both developed primarily after the Electronic Communications Privacy Act,[376] disrupt this

370. *See supra* pp. 32–34; *supra* pp. 58–59.

371. *See* 18 U.S.C. § 1030(a)(2)(C); *supra* pp. 74–75.

372. *See, e.g.,* Paul Belonick, *Transparency is the New Privacy: Blockchain's Challenge for the Fourth Amendment*, 23 Stan. Tech. L. Rev. 114, 153 (2020) (discussing, in the context of blockchain and the Fourth Amendment, whether digital signatures are content) (citing Riana Pfefferkorn, *Everything Radiates: Does the Fourth Amendment Regulate Side-Channel Cryptanalysis?*, Conn. L. Rev. 1429–30 (2017)). *Compare* Salgado, *supra* note 360, at 42 (arguing, in the context of hard drive searches, that hash comparisons are not Fourth Amendment searches because "the hash value is no more useful than a random number"), *and* Lin, *supra* note 360, at 1118–19 (same), *with* Martin, *supra* note 361, at 726–27 (arguing that hash-based screening of emails would "approximate the use of general warrants" disallowed under the Fourth Amendment), Kassotis, *supra* note 361, at 1313–14.

373. *See* Electronic Communications Privacy Act of 1986, Pub. L. No. 99-508, 100 Stat. 1848.

374. In 1974, the National Bureau of Standards proposed adopting a standardized algorithm for data encryption. *See* Encryption Algorithm for Computer Data Protection, 40 Fed. Reg. 12134 (Mar. 17, 1975).

375. *See* Electronic Communications Privacy Act of 1986, H.R. Rep. No. 99-647, at 37 (1986).

376. *See* Bart Preneel, *The First 30 Years of Cryptographic Hash Functions and the NIST*

logic. The information produced by both of these technologies is unintelligible on its own. Yet that information can be usefully intercepted, to authenticate a sender as tied to a message's content (in the case of a digitally signed hash) or to modify or alter the underlying plaintext message (in the case of homomorphic encryption). Because the Wiretap Act did not contemplate that unintelligible encrypted content could nevertheless have informational value, the statute offers no clear answers when applied to advanced cryptographic technologies that generate informational value from encrypted content.

How could the statutory schemes better accommodate these new cryptographic materials? One option would be to deem them neither content nor metadata, qualifying for protection under none of the statutes. This would have the benefit of avoiding legal liability for any of the discussed technologies, since these encrypted materials could be intercepted and used without restriction. This is probably not an ideal result, however. The longer encrypted data is retained, the more likely the underlying content will be revealed, either because advances in cryptanalysis break the encryption schemes over time or because the encryption keys fall into the hands of third parties.

Preferably, Congress would develop a statutory scheme specific to cryptographic hashes and other encrypted material. Such a statute would take into account both the technical needs to retain such encrypted material to effectuate content moderation, while limiting storage and distribution of such material in view of the risks of long-term retention. The statute would thus take an intermediate approach between § 2511's strict prohibitions on interception of content[377] and the SCA's and the PRA's permissiveness toward platform collection and use of metadata.[378]

2. Consent and Authorization

The interaction between consent and encryption provides another source of uncertainty. The Wiretap Act, the SCA, and the PRA provide

SHA-3 Competition, 2010 Topics Cryptology—CT-RSA 1, 4 (discussing development of early hash functions in the late 1980s and early 1990s); Craig Gentry, *Fully Homomorphic Encryption Using Ideal Lattices*, 41 Proc. Ann. ACM Symposium on Theory Comput. 169, 169 (2009) (proposing the first fully homomorphic encryption scheme).

377. *See* Wiretap Act, 18 U.S.C. (a).

378. *See* § 2702(a)(3); § 3121(b)(1).

exceptions based on user consent,[379] and the CFAA turns on "authorization."[380] Judicial decisions suggest different approaches for consent across the laws: Courts are reluctant to infer consent under the Wiretap Act absent a strong factual showing,[381] while implied consent under the CFAA appears to be found with some regularity.[382]

As identified in the discussions of those provisions,[383] these consent provisions highlight a tension with perhaps any content moderation technology applied to an end-to-end encrypted system. On the one hand, a platform's terms of service can presumably authorize the platform to use that technology, and the technology can serve important trust and safety objectives. On the other hand, a user's decision to use an end-to-end encrypted platform implies an intention to disallow the platform from accessing the user's messages or content; finding consent to the platform's content moderation seems to conflict with that explicit disallowance.[384]

Settling the requirements for consent across the communication privacy laws would of course help to reduce uncertainty and clarify the permissibility of content moderation technologies that work around end-to-end encryption. But the tension involved in the consent analyses highlights a broader question with the scope of the privacy expectations that end-to-end encryption entails. We explore that broader question below.[385]

3. Permitted Business Activities

Of the laws reviewed, the PRA was the only one that specifically addressed a platform's efforts toward "protection of users . . . from abuse of service."[386] That the communication privacy laws generally do not con-

379. *See* Wiretap Act § 2511(2)(d); 18 U.S.C. § 2702(b)(3); *id.* § 3121(b)(3).

380. *See* 18 U.S.C. § 1030(a)(2).

381. *See supra* pp. 34–35.

382. *See supra* note 211.

383. *See supra* pp. 34–35 (the Wiretap Act); *supra* pp. 75–76 (the CFAA).

384. *Cf. In re* Google Inc. Cookie Placement Consumer Priv., 806 F.3d 125, 151 (3d Cir. 2015) (holding that users' adoption of cookie blockers "clearly communicated denial of consent"); Grimmelmann, *supra* note 210, at 48–49 (noting difficulty in finding consent based on software terms of use, where software's activities conflict with other software the user has installed).

385. *See infra* Section IV.C.

386. *See* 18 U.S.C. § 3121(b)(1).

template content moderation is unsurprising for telephone-era statutes directed to one-on-one communications. Electronic group messaging capabilities, however, create opportunities for harassment, misinformation, spread of CSAM, and other forms of abuse. There is a growing perception that platforms should have an ethical duty, if not a legal one, to moderate content.

Although this Article has focused on content moderation on end-to-end encrypted systems specifically, it has highlighted a general need to clarify whether and when content moderation runs afoul of the communication privacy laws. Legislatively this could be achieved by adopting the abuse-protection exceptions from the PRA into the Wiretap Act and the SCA. But courts could achieve a similar result by clarifying that a platform's content moderation activities are within the "ordinary course of its business"[387] and "necessarily incident to the rendition of the service."[388]

It is not clear, though, that this is the best approach for adapting the communication privacy laws to platform content moderation. Platforms often moderate content for a variety of reasons unrelated to abuse protection: promoting diversity, balancing debate viewpoints, or responding to developing emergencies, for example. From that perspective, even the exceptions in the PRA may turn out to be undesirably narrow. More importantly, if platform liability turns on whether a certain content moderation practice falls within a statutory exception, that interposes the judiciary in setting platform content moderation policies, a traditionally private matter.

4. Computer Devices

Although the communication privacy laws typically apply to activities on a computer or electronic device, it is often not well-defined which devices fall within their contemplation, opening the door to some creative interpretations of the statutes. The Wiretap Act proscribes the use of a device to "intercept" communications, suggesting that the device should exist somewhere between the communicating parties and collect data in transit,[389] but parties have alleged, with some success, violations based on the

387. Wiretap Act, 18 U.S.C. § 2510(5).

388. § 2702(b)(3).

389. *See* Wiretap Act § 2510(4).

parties' own device and collection of data before or after message transit.[390] Similarly, SCA litigants have sometimes alleged that a personal computer is a "facility" protected from unauthorized access under the statute.[391] And the term "computer" in the CFAA, though seemingly referring to a single device,[392] could be interpreted to encompass an entire network of computers.[393]

At best, these creative interpretations of computer devices force courts into contorted efforts to twist the other elements of the statutory language to fit the theory of liability.[394] But they can also give rise to unexpected pathways to liability. The network trespass theory of the the CFAA, for example, was originally conceived as an argument to enhance platforms' ability to police problematic network behavior through legal action against malicious users.[395] Yet the same theory potentially limits platforms' ability to police problematic behavior through content moderation technologies, because treating an end-to-end encrypted messaging network as a single "computer" under the CFAA might render content moderation activities to be unauthorized access to that "computer."[396] Clarifying the scope of computer devices across the statutes would help to avoid interpretive difficulties arising from unconventional theories of what constitutes a relevant computer.

5. Statutory Modularity

The statutory concepts of content, authorization, computer devices, and business uses are largely common to all of the communication privacy laws. Yet each statutory scheme introduces its own definitions and exceptions to those terms, leaving each statute with idiosyncratic and in-

390. *See In re* iPhone Application Litig., 844 F. Supp. 2d 1040, 1062 (N.D. Cal. 2012); *In re* Pharmatrak, Inc. Priv. Litig., 329 F.3d 9, 22 (1st Cir. 2003); *In re* Google Inc. Cookie Placement Consumer Priv., 806 F.3d 125, 135 (3d Cir. 2015); *supra* pp. 35–36.

391. *See supra* note 186 (discussing cases).

392. *See* 18 U.S.C. § 1030(e)(1) (defining "computer" as "an electronic, magnetic, optical, electrochemical, or other high speed data processing device").

393. *See* Penney & Schneier, *supra* note 108, at 494–95.

394. *See, e.g., iPhone*, 844 F. Supp. 2d at 1058 (reasoning that if a user's device is a "facility" under the SCA, then the communications provider is a "user").

395. *See* Penney & Schneier, *supra* note 108, at 478–79.

396. *See supra* Section III.A.5.

consistent definitions. This is perhaps most noticeable with regard to the business use exceptions. The Wiretap Act and the SCA except activity that is "a necessary incident to the rendition of [a communication provider's] service or to the protection of the rights or property of the provider of that service."[397] The PRA instead excepts a range of activities relating to "operation, maintenance, and testing" of a service, "protection of the rights or property of such service," or protection "from abuse of service or unlawful use."[398] The CFAA has no business use exception, perhaps because it was assumed that a platform always had authorization to obtain content it handled.

Interestingly, most of the statutes do use a consistent definition of content. This is because the SCA, the PRA, and CALEA all incorporate the Wiretap Act's definitions by reference.[399] Indeed, the PRA specifically states that the scope of its coverage "shall not include the contents of any communication,"[400] neatly relying on the Wiretap Act's definition to carve up coverage between the two laws.

That model of consistency could be followed for the other statutory concepts identified above. For example, a single definition of acceptable business uses could be incorporated into all of the communication privacy laws, simplifying interpretation and avoiding the need to study each statute individually to discover one's legal obligations.

To be sure, there may be situations where divergent definitions are desirable. The CFAA, for example, likely prohibits unauthorized obtaining of "information" rather than "contents" because the statute is intended to proscribe unauthorized metadata capture. Nevertheless, having a single baseline definition of contents and metadata would still be helpful, as it would give legislators a unified set of statutory terms for defining "information" in the CFAA.

397. Wiretap Act, 18 U.S.C. § 2510(2)(a); Stored Commc'ns Act ("SCA").

398. *See* 18 U.S.C. § 3121(b).

399. *See* § 2711(1); § 3127(1); Communications Assistance for Law Enforcement Act (CALEA) § 102(1), 47 U.S.C. § 1002.

400. 18 U.S.C. § 3127(3)–(4).

6. CALEA Encryption Exception

Aside from general concerns about mandated design of technical systems, CALEA presents two lines of concerns with respect to its application to the content moderation technologies discussed. First, it could result in encrypted materials being retained for longer than it would be safe to do so.[401] Second, the privacy guarantees of technologies like message franking depend on separation of information between the platform and messaging users,[402] and technical design requirements under CALEA could vitiate that separation.

One possible way of addressing these problems would be to expand CALEA's existing exception for encryption.[403] While that exception currently provides that platforms "shall not be responsible for decrypting" communications, it could further absolve platforms of requirements to intercept encrypted materials in the first place. Platforms would still retain encrypted information such as traceback records in accordance with their content moderation needs, and law enforcement would essentially enjoy the same privileges to investigate encrypted communications as the platform would enjoy to moderate those communications.

B. Insights Into the Technologies

Our legal analysis of content moderation technologies also engages with a conversation about the technologies themselves. That conversation has already begun: The developers of these technologies have noted uncertainty about their own works' normative desirability,[404] commenta-

401. *See* CALEA § 103(a)(1) (requiring platforms to enable government interception of communications "at such later time as may be acceptable to the government").

402. *See supra* Section III.A.

403. *See* CALEA § 103(b)(3).

404. *See, e.g.*, Kulshrestha & Mayer, *supra* note 43, at 905 ("We do not take a position on whether E2EE services should implement the protocols that we propose, and we have both technical and non-technical reservations ourselves."); Tyagi et al., *supra* note 225, at 423 ("Robust policy dictating how and when to perform [message forward] tracing is necessary for protection of users' privacy expectations."); Issa et al., *supra* note 148, at 2337 ("[A]ny decision to use content moderation within end-to-end encrypted messengers requires weighing all of its potential benefits and risks. . . . We take no stance on the policy question . . . .").

tors have debated the human rights implications of these technologies,[405] and lawmakers have even introduced legislative and policy proposals on content moderation for end-to-end encrypted platforms.[406] But by systematically reviewing the legal elements of communication privacy and the technological elements of computer systems, we provide sharper focus on specific normative questions to be addressed. Insofar as the communication privacy laws are intended to reflect federal policy on reasonable expectations of privacy, tensions between the law and the technologies may point to larger societal tensions with these technologies.

The discussion of the business use exception under the Wiretap Act exemplifies this focus-sharpening.[407] The case law on telephone monitoring of employee conversations[408] suggests a distinction in legal treatment. Systems like traceback that retain content for later use are less likely to avoid liability based on the statutory exception,[409] while systems like server-side scanning that cryptographically manipulate content without retaining it are more likely to fall within the exception.[410] That difference in the law perhaps reflects a policy preference for data minimization, a preference that in turn can inform the future design of content moderation technologies.

Other findings from our legal analysis similarly provide guidance for future technological development. The possibility that CALEA could enable law enforcement to gain access to stored message franking or forward tracing information is a useful reminder that governments can influence

405. *See, e.g.*, Rosenzweig, *supra* note 303 (considering client-side scanning); BUS. FOR SOC. RESP., *supra* note 13 (Facebook-commissioned report discussing human rights implications of client-side scanning and other content moderation technologies for end-to-end encrypted platforms).

406. *See* Natasha Lomas, *UK Wants to Force Encrypted Platforms to Do CSAM-Scanning*, TECHCRUNCH (July 6, 2022), https://techcrunch.com/2022/07/06/uk-osb-csam-scanning/ (describing U.K. legislative efforts to require client-side scanning); Robert Gorwa, *European Security Officials Double down on Automated Moderation and Client-Side Scanning*, LAWFARE (June 15, 2022), https://www.lawfareblog.com/european-security-officials-double-down-automated-moderation-and-client-side-scanning (describing similar E.U. efforts).

407. *See supra* pp. 36–38.

408. *See* Deal v. Spears, 980 F.2d 1153, 1158 (8th Cir. 1992), *discussed at supra* pp. 37–38.

409. *See supra* pp. 51–52.

410. *See supra* p. 59.

the content moderation process, a possibility that already is driving some computer science research.[411] Uncertainty about whether cryptographic hashes are "content" under the Wiretap Act and the SCA[412] is consistent with many computer scientists' skepticism of whether hashes can be revealed to platforms without violating users' privacy expectations on end-to-end encrypted platforms.

C. What Is End-to-End Encryption?

A larger lesson that arises from our legal analysis is that these technologies challenge the notion of what end-to-end encryption is in the first place. At a surface level, end-to-end encryption could be defined as a system in which a message remains encrypted all the way to its destination.[413] More rigorous definitions expand on the guarantees that an end-to-end encrypted system makes, often focusing on confidentiality (unauthorized third parties cannot read messages), integrity (third parties cannot change message content), and authenticity (third parties cannot send messages purporting to be from others).[414] These privacy guarantees are not made to computer systems as a technical matter, but to users, or "ends," as a matter of system design.[415]

The contemporary debate over end-to-end encryption has treated these privacy guarantees as a binary matter, either intact or broken, leading to the seemingly irreconcilable positions between encryption advocates and

411. *See, e.g.,* Scheffler et al., *supra* note 348.

412. *See supra* pp. 32–34.

413. *See* Wong, *supra* note 11, § 10.1 (defining end-to-end instant message encryption as "a concept of securing communications between two (or more) participants across an adversarial path").

414. *See, e.g.,* Mallory Knodel et al., *Definition of End-to-End Encryption* 5 (Internet Eng'g Task Force, Internet-Draft draft-knodel-e2ee-definition-10, Apr. 20, 2023), https://datatracker.ietf.org/doc/draft-knodel-e2ee-definition/10/. There are differing views as to the precise list of guarantees. *See, e.g.,* Alec Muffett, *A Duck Test for End-to-End Secure Messaging* 7–8 (Internet Eng'g Task Force, Internet-Draft draft-muffett-end-to-end-secure-messaging-03, July 12, 2021), https://datatracker.ietf.org/doc/draft-muffett-end-to-end-secure-messaging/03/.

415. *See* Britta Hale & Chelsea Komlo, *On End-to-End Encryption* 6–7 (Cryptology ePrint Archive, Paper 2022/449, 2022), https://eprint.iacr.org/2022/449 (analyzing the concept of "endness"); Knodel et al., *supra* note 414, at 3; Muffett, *supra* note 414, at 11.

critics.[416] Indeed, platforms' content moderation activities on end-to-end encrypted messaging systems have already spawned debates over whether the platforms "break end-to-end encryption."[417] Yet the technologies explored in this Article show that the privacy guarantees of end-to-end encryption can be altered in more subtle, indirect ways.

Forward tracing, for example, can expose the identity of a message's sender in limited circumstances.[418] That is an alteration of the property of "deniability," the idea that senders of end-to-end encrypted messages cannot later be provably tied to their message content.[419] It is not clear whether deniability is one of the privacy guarantees of end-to-end encryption.[420] Furthermore, forward tracing alters the deniability guarantee in only a limited way: The basic implementation allows the platform alone to discover the message sender;[421] more advanced protocols impose even stronger limits on when deniability can be overcome.[422]

The computer science literature has sought to taxonomize and characterize how content moderation technologies alter the privacy guarantees of end-to-end encryption. Sarah Scheffler and Jonathan Mayer, for example, distinguish between "full client privacy" technologies and "partial client privacy" ones, the latter of which offer less privacy to senders of content

---

416. *See supra* Section I.B.

417. *See, e.g.*, Peter Elkind et al., *How Facebook Undermines Privacy Protections for Its 2 Billion WhatsApp Users*, ProPublica (Sept. 7, 2021), https://www.propublica.org/article/how-facebook-undermines-privacy-protections-for-its-2-billion-whatsapp-users (adding clarification to article on WhatsApp's content moderation practices, to note that moderation of user-reported messages does not break encryption); Whitney Kimball, *WhatsApp Moderators Can Read Your Messages*, Gizmodo (Sept. 7, 2021), https://gizmodo.com/whatsapp-moderators-can-read-your-messages-1847629241 (observing "a lot of confusion about what the [Facebook] means when it says 'end-to-end encryption' " in view of Facebook's moderation of WhatsApp messages); Abelson et al., *supra* note 13 (challenging client-side scanning technology).

418. *See supra* Section III.B.

419. *See supra* pp. 28–29.

420. *See* Knodel et al., *supra* note 414, at 5–6 (characterizing deniability as an "optional/desirable" feature).

421. This is because the platform alone has access to the full set of encrypted pointers with respect to traceback protocols, or the platform's secret key with respect to source tracking. *See* Tyagi et al., *supra* note 225, at 417; Peale et al., *supra* note 233, at 1487.

422. *See supra* pp. 50–51 and notes.

deemed illicit.[423] However, taxonomizing the alterations does not answer the normative question of which alterations "break" end-to-end encryption and which are acceptable or *de minimis*. As a result, computer scientists have reached differing conclusions on whether particular content moderation technologies are compatible with end-to-end encryption.[424]

As courts, policymakers, and legal commentators assess the legality and desirability of the burgeoning range of content moderation technologies for end-to-end encrypted platforms, they will have to decide what guarantees of privacy such encryption entails—they will have to say what end-to-end encryption is. The consent and authorization provisions of the communication privacy laws offer one possible place where the law may evaluate this question, at least to the extent that a court takes into account normative considerations to determine consent.[425] Laws on false or deceptive advertising are another place where law may weigh in on the definition of end-to-end encryption. And generally, the ongoing debate over end-to-end encryption will need to treat such encryption not as a binary matter, but as a spectrum of privacy guarantees with subtle variations enabled by modern cryptographic content moderation technologies.

Conclusion

Encryption has costs, but perhaps those costs are not as severe as they have seemed to be. This, at least, is the upshot of the computer-science research on content moderation in end-to-end encrypted media. We believe that legal scholars and policymakers need to take this computer-science research seriously; it reorients existing debates and opens up new lines of research. And we believe that computer-science researchers need to take

423. *See* Scheffler & Mayer, *supra* note 276, at 7 tbl.2 (characterizing literature into these categories). Scheffler and Mayer also provide other distinctions among content moderation technologies, such as server privacy and transparency. *See id.* at 4–5.

424. *Compare* Hale & Komlo, *supra* note 415, at 14 ("Message franking intuitively meets our definition of end-to-end encryption, because users voluntarily reveal specific messages sent to them to the service provider."), *with* Scheffler & Mayer, *supra* note 276, at 415 ("Under the proposed designs of message franking the E2EE deniability property will no longer hold against the moderator . . . ."). The discrepancy here arises in part because Hale and Komlo do not treat authentication as a guarantee of end-to-end encryption. *See* Hale & Komlo, *supra* note 415, at 12.

425. *See supra* Section IV.A.2.

the governing law seriously; it shapes what these systems can and cannot legally do. This Article is an exercise in taking both the "computer science" and "law" parts of "computer science and law" seriously.

On the one hand, the technical details matter. End-to-end encryption is not just a black box that makes content moderation impossible. The abuse-prevention protocols we have discussed enable very specific forms of detection and reporting, and they do not fit conveniently into existing statutory boxes. Arguments over encryption regulation must engage with this detail.

On the other hand, the legal details also matter. Technologists working on encryption schemes that facilitate content moderation must navigate a surprisingly complicated legal thicket. We have seen, for example, that although the Wiretap Act, the SCA, and the PRA are broadly parallel, their statutory exceptions diverge significantly when applied to encrypted content moderation. Real-world encryption systems will have to fit within these exceptions.