How To Think About End-To-End Encryption and AI: Training, Inference, Disclosure, and Consent

Mallory Knodel[1], Andrés Fábrega[2], Daniella Ferrari[1], Jacob Leiken[1], Betty Li Hou[1], Derek Yen[1], Sam de Alfaro[1], Kyunghyun Cho[1], and Sunoo Park[1]

[1]New York University
[2]Cornell University

In an era where so many forms of communication have moved online, widespread end-to-end encryption (E2EE) has become an essential underpinning for the security of communications, providing critical protections for journalists, dissidents, and activists, as well as the everyday family and social communications of billions. The movement towards widespread E2EE communication stayed strong alongside the "big data" trends of recent decades: a recognition of the value of keeping private messaging data out of increasingly vast repositories of personal information. Following remarkable recent advances and an explosion of interest in large language models and generative artificial intelligence (AI) more broadly, however, we observe three trends that raise alarms for E2EE security.

First, the way people interact with AI models is changing. While they originally served as standalone applications, AI models are now increasingly incorporated into other everyday applications and throughout devices, including messaging applications, in the form of AI "assistants." Interacting with these assistants is often baked into the user experience by default, made readily available as part of the application client (e.g., within a messaging app).

Second, high-quality training data is becoming scarce. This has created a race between model developers, with tech companies under increasing pressure to tap any potential source of human-written content. That which is publicly accessible online has already been harnessed, so privately held data is naturally the next resort, and indeed, companies have been quietly changing their terms of service to enable training AI models on more and more of the user data they hold [3].

Third, recent trends in AI appear to have shifted business incentives. Over the last decade, E2EE became a central part of the business model of online intermediaries. Business incentives to protect proprietary and user data were aligned with human rights protections. However, in the last couple of years we have seen a shift towards prioritizing AI features as a component of modern applications, and a corresponding search for sources of data to power these, sometimes even at the cost of existing product features and profits [2, 1].

Against the backdrop of these three concerning trends, the motivating question of our talk is as follows:

Is processing of E2EE content by integrated AI models compatible with end-to-end encryption?
(If so, to what extent and under what circumstances?)

Summary of contributions our talk will cover. We systematically analyze the above question from both a technological and a legal perspective, combining the expertise of an multidisciplinary research team across cryptography and security, AI and LLMs, and law (touching on contract law, data protection law, consumer protection law, and antitrust).

We identify the key confidentiality and integrity properties provided by E2EE, for both *individual* and *systemic* privacy considerations, which might be adversely impacted by feeding E2EE content into AI models. This requires a novel conceptualization of key properties of E2EE, ranging from the strictly definitional, to essential properties that only hold when E2EE is used at scale, to many important properties that are widely associated with E2EE but not technically within the scope of standard academic definitions.

We then examine a wide range of detailed technical configurations that could fall under the broad umbrella of "feeding E2EE content to AI models," taking into consideration the state of the art in cryptography and

privacy technologies, as well as the latest developments in AI/ML. We distinguish the use of E2EE content for *inference*, for *training*, or for both. We assess the capacity of each technical configuration—that is, of each distinct approach to adding AI features to an initially E2EE system—to uphold the key guarantees of E2EE. We conclude that some configurations *cannot* uphold these guarantees, while some others can.

Next, we overview potentially relevant areas of law, and provide a detailed analysis of the circumstances under which E2EE service providers are likely to be able to offer AI features which use E2EE content, consistent with their legal obligations under current US and EU law (alongside a brief note on other jurisdictions). We highlight areas of legal uncertainty, and provide a detailed exposition of pending legal (or quasi-legal) processes where relevant. In addition, looking beyond the limits of existing law, we discuss relevant critiques of current law and diverse academic theories of consent and privacy (legal and otherwise), and briefly explore the application of such critiques and theories to the use of E2EE data for AI.

Putting together all of the above, our analysis yields the following key recommendations. Key terms within the recommendations are indicated in italics; precise definitions of the key terms will be provided in the talk and in our paper.

1. **Training.** Using end-to-end encrypted content to train *shared AI models* is not compatible with E2EE.
2. **Inference.** Querying AI models with *E2EE content* may be compatible with end-to-end encryption only if the following recommendations are upheld:
   (a) Prioritize *endpoint-local* processing where possible.
   (b) If performing inference with *non-endpoint-local* models,
      (i) No *third party* can see or use any *plaintext-dependent* versions of *E2EE content*,[1] and
      (ii) A user's *E2EE content* is exclusively used to fulfill that user's queries.
3. **Disclosure.** Messaging providers should not make unqualified representations that they provide E2EE if the default for any conversation is that *E2EE content* is used (e.g., for AI inference or training) by any *third party*.
4. **Opt-in consent.** AI assistant features, if offered in E2EE systems, should generally be off by default and only activated via opt-in consent. Obtaining meaningful consent is complex, and requires careful consideration including but not limited to: scope and granularity of opt-in/out, ease and clarity of opt-in/out, group consent, and management of consent over time.

Our analysis indicates that there would be serious risks (both technical security risks and legal risks) involved in diverging from these recommendations and thus undermining the established operation of the current E2EE ecosystem according to privacy-by-default practices. The greatest risk is in eroding E2EE as a baseline systemic guarantee—an assumed condition of any application or service that is held to be E2EE—and turning it instead into merely a bonus feature that is easily and routinely compromised in exchange for convenience. This would compromise the "network effects" that currently protect privacy and freedom of expression for at-risk groups and everyday users around the world.

References

[1] Benj Edwards. *So far, AI hasn't been profitable for Big Tech.* `https://arstechnica.com/information-technology/2023/10/so-far-ai-hasnt-been-profitable-for-big-tech`.

[2] Tom Lewis. "How Generative AI Is Making Customer Experience Worse". In: *Forbes* (2024).

[3] Eli Tan. *When the Terms of Service Change to Make Way for A.I. Training.* `https://www.nytimes.com/2024/06/26/technology/terms-service-ai-training.html`. 2024.

[1] The talk will provide detailed discussion of technical configurations that may achieve this condition.