The Day the Provider Stopped Reading Your Chats

WhatsApp Introduces Encrypted Chats with Meta AI

Theodore Christakis & Peter Swire

Theodore Christakis is Professor of International, European and Digital Law at University Grenoble Alpes (France), Director of Research for Europe with the Cross-Border Data Forum, Member of the Board of Directors of the Future of Privacy Forum and a former Distinguished Visiting Fellow at the New York University Cybersecurity Centre. He is co-Chair with Héber Arcolezi on "Responsible AI: Design, Regulation and Conformity" at the Multidisciplinary Institute in Artificial Intelligence.

Peter Swire is the J.Z. Liang Chair in the Georgia Tech School of Cybersecurity and Privacy, and Professor in the Scheller College of Business. He is Senior Counsel with Alston & Bird LLP, and Research Director of the Cross-Border Data Forum. After the Snowden revelations, Swire served as one of five members of President Obama's Review Group on Intelligence and Communications Technology. Under President Clinton, Swire was the Chief Counsellor for Privacy, the first person to have U.S. government-wide responsibility for privacy policy

Cover Illustration:
Original illustration conceived and directed by the authors and produced with generative AI assistance.

To cite this study: T. Christakis, P. Swire, The Day the Provider Stopped Reading Your Chats: WhatsApp Introduces Encrypted Chats with Meta AI, AI Regulation Papers, 26-05-4, AI-Regulation.com, May 2026.

This work has been partially supported by MIAI @ Grenoble Alpes, (ANR-23-IACL-0006) and by the Interdisciplinary Project on Privacy (IPoP) of the Cybersecurity PEPR (ANR 22-PECY-0002 IPOP).

AI-REGULATION PAPERS · 2026-05-4 · MAY 2026

The Day the Provider Stopped Reading Your Chats

WhatsApp Introduces Encrypted Chats with Meta AI

Theodore Christakis and Peter Swire

ABSTRACT

On 13 May 2026, Meta announced *Incognito Chat with Meta AI*, the most significant structural move in consumer chatbot privacy of the past two years. With this announcement, the architecture that Part 1 of Christakis's study on consumer chatbot confidentiality has been calling Sealed Mode ceased to be a research aspiration and became a deployed mass-market product. Users of WhatsApp can now chat with Meta AI in a mode in which the provider cannot read the conversation, in the closest functional equivalent to the way Meta cannot read WhatsApp messages between two users. The protection is architectural rather than contractual: the chat is processed inside a Trusted Execution Environment in which Meta has renounced, by hardware design, the capacity to access content.

This paper offers a first analytical reading of the announcement. We examine the cryptographic architecture and the external audits that support the trust claims; why this is best understood as a Sealed Mode deployment in the embedded-in-E2EE environment that Part 2 had identified as the most architecturally exposed; the moderation question ("cannot read, cannot moderate" on the provider side, with automated moderation continuing "inside the bubble"); the legal-architectural consequence ("no knowledge, no liability"); the new dimension the announcement opens in the long-running Going Dark debate; the constraints the architecture imposes on training and conversational memory; and the questions that the announcement leaves open, including whether the standalone chatbot providers (OpenAI, Anthropic, Google, xAI) might follow Meta's lead for at least some of their services.

On 13 May 2026, Meta announced *Incognito Chat with Meta AI*, the most significant structural move in consumer chatbot privacy of the past two years. It is also, on our reading, the first real-world deployment of the kind of architecture that Christakis's Part 1 of his study on consumer chatbot confidentiality has been calling **Sealed Mode**: a category of conversation in which the provider verifiably narrows its own access, by architecture rather than by policy, from the moment of collection. What follows is a first analytical reading of the announcement, with cross-references to Christakis's Part 1 (published March 2026), to the forthcoming Part 2 (May 2026), and to Swire, Ahmad and Specter's forthcoming work explaining end-to-end encryption (E2EE) for a law and policy audience. We are posting this promptly after the official announcement, to assist public understanding of the topic, but have not had the opportunity to read other public comments or have our text reviewed by other readers. We address questions the architecture raises about safety, liability, regulatory politics, and the price the user pays for protection, and

we conclude with the question of whether the same architecture can and should now be deployed by the standalone consumer chatbot providers.

1. What has been announced

For roughly the past year, Meta has been deploying a system called Private Processing for specific AI features inside WhatsApp, namely message summarisation and writing assistance.[1] What makes last week's announcement potentially significant is that the same architecture is being extended to the chats themselves between a user and Meta AI in WhatsApp. Until now, a user who summoned Meta AI inside WhatsApp (via the @Meta AI command or a dedicated icon) had their prompt and the AI's response leave the end-to-end encrypted pathway and reach Meta's servers in a form readable to Meta's infrastructure, in the same way that any cloud chatbot conversation to date reaches the provider in a form the provider can read.[2] That is the architectural anomaly the announcement is meant to close.

Meta has branded the new product *Incognito Chat with Meta AI*, and is launching it on WhatsApp and the standalone Meta AI app, with rollout expected over the coming months.[3] Three features of the design that Meta and WhatsApp emphasise in their public announcements are worth flagging at the outset, as they reflect the substantive guarantees the company is making. First, the chats are protected by encryption and the Trusted Execution Environment that Meta calls Private Processing, such that no one, including Meta and WhatsApp themselves, can read them. Second, the chats are not saved or stored: they disappear once the user exits the chat, and Meta AI loses the context of the conversation. Third, web search is also conducted privately: when Meta AI consults a search engine to obtain up-to-date information, the search is not linked to the user's identity, and users may also disable web search entirely. The feature is text-only (no image upload or generation). Meta has also pre-announced *Side Chat*, a forthcoming feature also built on Private Processing, that will allow Meta AI to be invoked within existing WhatsApp conversations to give private help with context of what is being discussed, without disrupting the main conversation.

The architecture rests on three components. The first is the **Trusted Execution Environment**, or TEE: a hardware-isolated zone inside a server, in this case built on AMD's SEV-SNP chips and Nvidia's confidential GPU platforms. Inside this zone, data can be decrypted and processed but is designed to remain inaccessible to the operating system on which the server runs, to the hypervisor that manages it, and to Meta's own administrators.[4] In ordinary cloud computing the question is whom you trust to refrain from reading; in confidential computing the claim is that no one can read, because the hardware itself enforces the boundary.[5] The second component is **Oblivious HTTP**, or OHTTP, a routing layer in which the user's device sends the encrypted request through an independent third-party relay (in this case Fastly) before reaching Meta's

[1] Meta Engineering Blog, *Building Private Processing for AI Tools on WhatsApp*, 29 April 2025.
[2] Electronic Frontier Foundation, *What WhatsApp's "Advanced Chat Privacy" Really Does*, 26 September 2025. The EFF analysis is particularly useful for distinguishing what Advanced Chat Privacy does and does not do, and for confirming that user-to-Meta-AI exchanges, prior to last week's announcement, did not enjoy end-to-end encryption.
[3] WhatsApp Blog, *Introducing Incognito Chat with Meta AI: A completely private way to chat with AI*, 13 May 2026; Meta Newsroom, *Introducing a Completely Private Way to Chat With AI*, 13 May 2026; WhatsApp Help Center, *How Incognito Chat with Meta AI works*.
[4] Meta AI, *Private Processing for WhatsApp: Technical Whitepaper*, v2 updated 16 March 2026.
[5] We have not done a technical assessment of the quality of Meta's implementation of a TEE. Researchers including Daniel Genkin have previously found vulnerabilities in TEEs. For example, see Chuang, Seto et al., *TEE.fail: Breaking Trusted Execution Environments via DDR5 Memory Bus Interposition* (2026); *Year of vulnerability hunting uncovers potential attacks on Intel Chips, RAM* (2019).

infrastructure. The relay sees who the user is, but not what the user is sending; Meta's gateway sees the encrypted payload, but not the user's identity. Neither party, on its own, has the complete picture.[6] The third component is **remote attestation**: before the user's device sends any content, it asks the server's hardware to produce a cryptographic "quote" certifying that the code running inside the TEE is the exact version that Meta has publicly logged and that independent auditors have reviewed. If the quote does not match, the device refuses to send the data.

The cryptographic workflow is therefore designed to operate as follows. The user's device asks Meta AI a question. The device contacts a third-party relay, which forwards the request to Meta's gateway without revealing the user's identity. The device receives the hardware's attestation quote, verifies it, and then encrypts the actual content with an ephemeral key shared only with the TEE. Meta's infrastructure sees only the encrypted payload pass through; the operators cannot decrypt it. Inside the TEE, the message is decrypted, the language model (Muse Spark in the current configuration) generates a response, the response is re-encrypted with the ephemeral key, and only the encrypted response leaves the enclave. When the request is complete, the TEE purges the data from memory. Nothing is written to disk. Nothing persists.

Two external security audits provide some grounds for treating these claims as more than marketing. Trail of Bits published a 28-issue review in early 2026 that identified eight high-severity findings affecting the integrity of the enclave (including a path by which a malicious hypervisor could have used fake ACPI tables to read protected memory). All eight were remediated prior to production.[7] NCC Group conducted a 115 person-day assessment of the message-summarisation service in late 2025, confirmed the statelessness claim on the basis of source-code review, and identified the residual risk that the OHTTP anonymity guarantee depends on Meta not colluding with the third-party relay.[8] The audits do not establish that the system is unbreakable;[9] they establish that the architecture is real, that the trust assumptions are limited and externally documented, and that the major attack surfaces have been examined by independent specialists.

2. Is it really equivalent to end-to-end encryption?

The honest answer is: it is not the same thing as classical end-to-end encryption, but it is the closest functional equivalent currently available for a service that requires server-side computation. The distinction matters and is worth being precise about. As Swire, Ahmad and Specter set out in their forthcoming work explaining E2EE for a law and policy audience, what counts as end-to-end encryption in any given system depends on the analytical question of which parties are treated as "ends", an inquiry the authors trace to the work of Chelsea Komlo and Britta Hale on the concept of *endness*.[10] The Meta announcement is best understood as a deliberate redefinition of where the "end" of the encryption sits in a system that also performs server-side computation.

---

[6] Cybernews, *WhatsApp announces Private Processing so users can use AI and preserve their privacy*, 2 May 2025.

[7] Trail of Bits Blog, *What We Learned About TEE Security from Auditing WhatsApp's Private Inference*, 7 April 2026. The audit identified 28 issues including 8 high-severity, all of which Meta remediated before deployment.

[8] NCC Group, *Public Report: Meta WhatsApp Message Summarization Service*, August 2025.

[9] Even for systems that have encryption at each stage of the process, "side channel attacks" may exist, where there may be leakage of data from a physical cryptosystem. See Hassan, Roy et al., *Memory Under Siege: A Comprehensive Survey of Side-Channel Attacks on Memory*, 8 May 2025.

[10] Peter Swire, Kenesa Ahmad & Michael Specter, *Encryption and Globalization 15 Years Later: End-to-End Encryption and the Third Round of the "Going Dark" Debate*, forthcoming, *Fordham International Law Journal*, 2026.

In the classical E2EE model, as deployed in messaging between two human users on WhatsApp or Signal, the provider's servers never possess the decryption keys. The message exists in plaintext only on the sending device and the receiving device. The provider routes ciphertext and nothing more. This model works because the recipient is a human user with a device that can perform the decryption locally. It does not work when the recipient is a large language model that requires server-side GPU resources too computationally heavy to run on a phone.

Private Processing addresses this by relocating the "endpoint" of the encryption from the recipient's device to a hardware-isolated zone inside the provider's server. The provider as an institution, namely its employees, administrators, ordinary servers, training pipelines, advertising systems, does not have the keys; only the TEE does, and only for the duration of the request. The conversation is end-to-end encrypted between the user's device and the TEE, but the TEE is physically inside Meta's data centre. The protection is therefore architectural rather than absolute: it depends on the integrity of the hardware enclave, the correctness of the code running inside it, and the trustworthiness of the chip manufacturer and of the third-party relay. None of these dependencies exist in the classical user-to-user encryption model. All of these dependencies are real and have been the subject of independent audit work.

A useful way to put the point is that Private Processing produces what one might call **operator-blind processing**: Meta retains the computational role of the service but renounces, by architecture, the ability to read what the service computes. That is not the same as user-to-user E2EE, but it is genuinely different from the ordinary cloud-AI model in which the operator can, in principle, read everything. For a user choosing between a chatbot whose operator can read every conversation and is bound only by terms of service, on the one hand, and a chatbot whose operator has architecturally renounced that ability, on the other, the difference is real.

3. Sealed Mode, from concept to deployment

In Part 1 of his study on consumer chatbot confidentiality, published in March 2026, co-author Christakis proposed an architectural concept called **Sealed Mode**.[11] The concept emerged from an examination of how the five most prominent consumer chatbot providers (ChatGPT, Claude, Gemini, Grok, DeepSeek) actually treat the conversations users have with them across four established boundaries of confidentiality (training, human review, advertising, and operational sharing) and a fifth emerging boundary (persistent memory and longitudinal profiles). The empirical finding was that terms of service that reserve broad rights to the provider are not, on their own, a credible foundation for user trust where the conversations contain health, financial, legal or other intimate information. The proposal that followed was structural rather than contractual: a category of conversation in which the provider would *verifiably* narrow its own access, retention, training use, advertising signals and downstream reuse, from the moment of collection, by architecture rather than by policy promise.

The six elements of the Sealed Mode proposal were: no use for training; no human review except by narrowly defined exception with user consent; siloed personalisation (personalisation that remains within the sealed environment, not flowing to advertising, training, or external systems); no advertising signals derived from the content; strict retention limits; and cryptographic hardening calibrated to the deployment environment.

[11] Theodore Christakis, *You Trust Your Chatbot With Everything. Should You? Part 1: How the Controller Uses Your Chat Data*, AI-Regulation Papers, March 2026. The Sealed Mode proposal is developed in chapter 6 of that study.

Private Processing, on the public description of its architecture, instantiates four of the six Sealed Mode elements directly, takes a more restrictive position than Sealed Mode envisaged on a fifth, and instantiates the sixth (cryptographic hardening) more comprehensively than any deployed consumer chatbot architecture we are aware of. Specifically: Meta's whitepaper describes the TEE as providing ephemeral data processing, with no persistent storage and the data purged from memory at the end of each request, which satisfies the strict-retention element and which entails, as a structural consequence, that the content of Private Processing conversations cannot be used to train future models. Human review of the in-enclave content is architecturally impossible without explicit user action (the system is designed so that human review can take place only when users themselves choose to report a particular message). On advertising, Meta does not address the point explicitly in its public documentation, but the conclusion follows from the architecture: content the provider cannot read cannot be used to derive advertising signals.[12]

On personalisation, Private Processing currently takes the most conservative position available: namely full statelessness, with no personalisation across sessions at all. This is more restrictive than the siloed-personalisation element of Sealed Mode, which envisaged personalisation that remains within the sealed environment but does not flow to advertising, training, or external systems. Meta has, in effect, chosen to forgo personalisation entirely in this iteration rather than to engineer a confidential-personalisation pathway. Whether the architecture will, over time, be extended to support some forms of confidential personalisation is one of the open questions to which we return below. For now, what should be observed is that the statelessness choice is more protective than the siloed-personalisation element of Sealed Mode required, at the cost of a meaningful degradation in the user experience.

The sixth Sealed Mode element, cryptographic hardening, is exactly what Private Processing's TEE-plus-attestation-plus-OHTTP stack is intended to provide.

We do not claim, and would not claim, that Meta's engineering team built Private Processing in response to Christakis's March 2026 study, which was published well after the Private Processing framework was first announced in April 2025. The intellectual lineage runs the other way: Private Processing and adjacent confidential-computing architectures, notably Apple's Private Cloud Compute announced in June 2024,[13] are the technical infrastructure that made it possible to propose Sealed Mode as more than a conceptual aspiration. What appears genuinely new about last week's announcement is the *scope* of the deployment. Until now, Private Processing has covered specific peripheral features (summarisation, writing help) applied to existing WhatsApp messages. Extending the same architecture to the chat conversation itself with Meta AI moves a major consumer chatbot interaction into something approaching Sealed Mode. That is a different kind of deployment from the partial application that has existed since 2025.

[12] The conclusion of no use of the plaintext for advertising assumes there is no mechanism where information about the plaintext leaves the TEE to be sent to a destination that uses the data for advertising. Based on available information, we believe that Meta is blocking such information from the TEE both to Meta and to any other recipient, such as an advertiser.

[13] Apple Security Research, *Private Cloud Compute: A New Frontier for AI Privacy in the Cloud*, June 2024. The Apple architecture is, in its essentials, the same pattern Meta has now adopted: TEE-based inference, hardware-attested code, no operator access to plaintext.

4. Why deployment environment matters

In Part 2 of his study, which is about to be published, Christakis has devoted an entire chapter to the tension between cryptographic hardening and content-based safety monitoring.[14] One of the central arguments of that chapter is that the appropriate level of cryptographic protection for a chatbot is not a single uniform answer for the category of "chatbots", but depends on the *deployment environment* in which the chatbot operates, because that environment carries prior architectural commitments that the chatbot inherits.

The argument distinguishes three deployment environments. The first is the **standalone consumer chatbot**, such as ChatGPT, Claude, Gemini, or Grok accessed through a dedicated interface, where no prior cryptographic commitment exists and the encryption question is an open product-design choice. The second is the **chatbot embedded in an end-to-end encrypted messaging environment**, such as Meta AI in WhatsApp, where the surrounding platform has made a deliberate prior commitment that the provider cannot read messages between users, and where any departure from that posture by the embedded chatbot materially weakens the platform's overall security stance. The third is the **enterprise or health-specific chatbot operating under Sealed Mode**, where the appropriate level of protection will depend on the specific regulatory and clinical context.

The point of the typology was that Meta AI embedded in WhatsApp falls into the second category, and that the architectural anomaly to be solved was precisely that the chatbot pathway in an otherwise end-to-end encrypted environment was leaving plaintext on Meta's servers, in apparent tension with the platform's prior commitments to its user base. The relevant passage from the forthcoming study reads: "Embedding an AI assistant into [an end-to-end encrypted] platform in a way that bypasses [the platform's prior architectural] choice (by routing queries in plaintext through provider servers, by retaining records on the provider side, or by enabling content-based safety monitoring of conversations that would otherwise be inaccessible to the provider) materially weakens the platform's overall security posture without necessarily being visible to its users."

The announcement made by Meta on 13th May, if its architecture matches the Private Processing description, is the response to this concern. Instead of leaving plaintext on the provider's servers, Meta's chatbot pathway now passes through a TEE-protected enclave. The departure from the surrounding platform's encryption posture is no longer absolute. The architectural coherence of WhatsApp as an end-to-end encrypted platform is, on this design, significantly more defensible than it was prior to last week's announcement.

5. The moderation concern: cannot read, cannot moderate

The architectural turn has a consequence that runs through the rest of the analysis and that deserves its own treatment. If Meta (or, in the encrypted-mode scenario described in the conclusions, any other provider) cannot read the contents of a conversation, then it also cannot moderate the contents of that conversation in the way that an ordinary cloud-AI provider does, by inspecting the input, applying policy criteria, and refusing or escalating problematic exchanges. The moderation pipeline that is used in consumer AI (automated classifiers, human reviewers, escalation channels, law-enforcement referrals where the threshold is met)

[14] Theodore Christakis, *You Trust Your Chatbot With Everything. Should You? Part 2: Governments, Courts, and the Battle Over Your Chatbot Logs*, AI-Regulation Papers, forthcoming May 2026.

presupposes that the provider has plaintext access to the conversation in the first place. Take that access away and the pipeline cannot operate in its familiar form.

Some architectural features enable nonetheless a moderation process, although they differ from the previous moderation pipeline.

In-enclave moderation operates through two distinct mechanisms inside the secure environment, both invisible to Meta's external systems. The first is **refusal at the model level**: through supervised fine-tuning and reinforcement learning from human feedback, the language model has been trained to refuse certain categories of request, so that a user who asks the chatbot how to build a bomb receives a refusal generated by the model itself.[15] The second is **classifier-based moderation**: alongside the language model, Meta runs safety classifiers in the TEE (Llama Guard 2, which monitors prompts and responses for harmful content; Llama Code Shield, which monitors for vulnerable code suggestions; CyberSec Eval 2, which assesses cyberattack-assistance propensity), which can trigger a refusal independently of the model's own behaviour if a defined threshold is crossed. In both cases the relevant computation is performed inside the enclave, with full plaintext access to the conversation; the refusal is generated and encrypted inside the enclave; and what leaves the enclave, to be transmitted back to the user, is an encrypted "I cannot help with that" response. Meta's external systems, namely the systems that handle ordinary logging, advertising signals, and content moderation review, never see the prompt that was refused, the classifier's judgement, or the refused response. Moderation happens, but it happens **inside the bubble**. The provider sets the threshold for triggering a refusal, but the actual refusals are invisible to the institution that runs the bubble.[16]

Furthermore, Meta, like all major chatbot providers, has other mitigations beyond simple refusal. For categories of harm such as suicide and self-harm, these models are generally trained to direct users to expert resources or support in their responses, rather than to refuse outright. For repeat attempts to bypass the safety layer, the user is temporarily blocked from further chatbot responses.[17] By the architecture of the TEE, neither mitigation requires the provider to see the content of the underlying conversation in order to function: both can operate entirely inside the enclave.

The combination is not a perfect substitute for content-based moderation by the provider. There are categories of harm that classifier-plus-refusal architectures will fail to catch, for the same reasons they fail to catch them in non-confidential deployments, such as: ambiguous requests, novel jailbreaks, fictional or roleplay framings that conceal genuine intent, multi-turn manipulations that build up to a harmful request across many exchanges. Confidential deployment does not solve the underlying classification problem; it only constrains where the classification happens and who can see the result. What it offers, in exchange for that constraint, is a meaningful and architectural reconfiguration of the trust relationship between the provider and the user.

6. No knowledge, no liability: a legal-architectural consequence

The architectural move has a legal-institutional consequence that has not yet, as far as we are aware, received the analytical attention it deserves. If a provider cannot read the contents of a

[15] The TEE architecture affects model-level moderation to the extent that the model no longer trains on content that is processed within the TEE.
[16] Meta AI, *Our responsible approach to Meta AI and Meta Llama 3*.
[17] See the forthcoming Part 2 of Christakis' study on these mitigation features.

conversation, then the provider also cannot, in the legally relevant sense, *know* their contents. And much of the liability framework that has been developing around consumer chatbot providers (in the United States particularly, but increasingly in the European Union as well) turns on what the provider knew or should have known.

Chapter 1 of the forthcoming Part 2 of Christakis's study sets out how recent litigation against chatbot providers all turn on theories of provider knowledge, including the Tumbler Ridge wrongful-death suits filed in California in April 2026 against OpenAI, the *Garcia* and *Raine* cases relating to minors' suicide, and the Florida criminal investigation of OpenAI announced by Attorney General Uthmeier on 21 April 2026.[18] The plaintiffs argue, in various forms, that the provider knew or should have known that its system could produce harmful outputs and that it failed to act on that knowledge. The chapter develops the structural point at length: the more rigorously a provider monitors its product for signs of imminent harm, the richer the evidentiary record a future prosecutor or civil plaintiff can use to support a knowledge-based claim against it. The perverse incentive structure that has been long observed in safe-harbour debates for online intermediaries is now becoming visible in the chatbot setting as well.

Private Processing, by design, reduces the actual knowledge of the provider. If the architecture works as advertised, Meta has, in the relevant sense, no internal records of what its users discussed with Meta AI through the encrypted-chat pathway, beyond the bare fact that a request was made. The provider has architecturally renounced not only the capacity to read but also the capacity to be later found to have known. That is a substantial reduction in the liability surface. Plaintiffs would apparently continue to be able to argue that the provider "should have known" enough to take action to reduce the harm. For that reason, the accountability framework, if it is to operate at all in this architecture, has to be relocated to the design phase, where the criteria embedded in the safety classifiers and the model's refusal patterns are decided, rather than to the operational phase, where individual conversations are reviewed.

This change in the provider's actual knowledge may result in reduced effectiveness for the kinds of provider-side accountability mechanisms that some of those involved in this debate (including Christakis) have been calling for in the proactive-disclosure context. Going forward, the number of referrals reported in transparency reports would presumably be lower, because the provider would less often have enough knowledge to trigger a referral. External review of how providers set referral criteria would similarly take place without access to the content processed in the TEE.

These effects on liability, we would suggest, may be significant, and may not be otherwise highlighted in the initial commentary. The architectural protections that Private Processing affords may thus run counter to some of the accountability mechanisms that have been supported by plaintiffs and those who agree with them.

7. Going Dark, in a new dimension

The architectural move will quite possibly raise concerns from law-enforcement agencies and organisations focused on the detection of online harms, including child sexual abuse material. The arguments will be familiar from the long debate about end-to-end encrypted messaging that

[18] The Tumbler Ridge wrongful-death and personal-injury suits filed in California federal court on 29 April 2026 against OpenAI and Sam Altman personally; the *Garcia v. Character Technologies* litigation (settled January 2026) and the Adam Raine litigation against OpenAI filed August 2025; and the Florida criminal investigation announced by Attorney General James Uthmeier on 21 April 2026 in connection with the April 2025 FSU shooting. All case-clusters are analysed in detail in the forthcoming study of Christakis, Part 2, chapter 1.

has run from the original Crypto Wars of the 1990s through the 2015–2016 FBI–Apple confrontation over the San Bernardino iPhone, the European debate around proposed CSAM-scanning regulation, and the periodic resurgence in the United States of legislative proposals that would condition intermediary liability protections on content-scanning practices. Last week's announcement is, in our reading, best understood as part of what Swire, Ahmad and Specter have called Round 3 of the Going Dark Debates, the period since 2015 in which E2EE has become the default for major categories of consumer communications and in which law-enforcement bodies have responded with renewed proposals for limits on its use.[19] Their analysis offers a useful frame for thinking about how the chatbot version of the debate is likely to unfold.

Private Processing extends the Going Dark debate to a new context, and the threshold question is whether the chatbot version is a genuinely new debate or the same debate in a new setting.

The case for treating it as new rests on the provider's role in the conversation. In the classical user-to-user setting, the provider is a routing intermediary between two human parties who have themselves chosen how to communicate. In the chatbot setting, the provider operates one end of the conversation: the AI is the provider's own product. Two consequences follow from this difference. First, the object of the law-enforcement demand changes. In the classical case, the state asks the provider to expose communications between third parties it merely routes; in the chatbot case, the state asks the provider to retain the capacity to read its users' interactions with its own chatbot. Second, the architecture creates a novel accountability gap: because the model itself runs inside the enclave and its outputs are encrypted before they leave, the provider originates content it cannot inspect, a gap that has no real analogue in user-to-user encryption, where the humans on both sides are the authors of what they exchange.

The case for treating the chatbot debate as structurally the same starts from a different angle. In both settings, the provider has made an architectural choice to encrypt, the user has made a choice to use the encrypted service, and unencrypted alternatives remain available in the market; the contested act in both cases is the corporate decision to render content unreadable to the provider, and the state's intervention takes the same form. On this reading, the chatbot debate is the latest iteration of a policy question running from the Crypto Wars of the 1990s to the present, and should be resolved on the same grounds. Whatever the structural differences between user-to-user and user-to-AI encryption, the question that decides the case is the same in both, and it is at root a proportionality question: whether the investigative benefits of compelling a provider to retain readability justify the privacy, cybersecurity, and other costs imposed on the overwhelming majority of users whose conversations are of no law-enforcement interest.

The empirical case for the law-enforcement side of that proportionality question has been the subject of sustained scrutiny in Swire, Ahmad and Specter's analysis of what they describe as three rounds of the Going Dark Debates. They have argued, across those rounds, that the case for "going dark" is weaker than law-enforcement framings typically suggest, notably because what they call the "Golden Age of Surveillance" has produced an unprecedented richness of investigative material accessible through other means: location records, device metadata, social-graph data, facial recognition, behavioural databases, and the ordinary content available through default cloud backup and software-as-a-service environments. That analysis applies to the

[19] Peter Swire, Kenesa Ahmad & Michael Specter, *Encryption and Globalization 15 Years Later: End-to-End Encryption and the Third Round of the "Going Dark" Debate*, forthcoming, *Fordham International Law Journal*, 2026. The article updates and extends the 2012 analysis by Swire and Ahmad of the relationship between encryption, lawful access, and the global Internet, and identifies what the authors call "Round 3" of the Going Dark Debates: the period since 2015 in which end-to-end encryption has become the default for large categories of consumer communications, including (with last week's Meta announcement) for at least one major consumer chatbot.

chatbot setting. The architectural inability to read a particular subset of conversations with Meta AI through Private Processing does not, in itself, constitute the kind of investigative blindness that law-enforcement framings sometimes describe. It constrains one particular evidentiary surface while leaving many others operative.

The Instagram counter-data point is worth noting in this context. On 8 May 2026, four days before Meta's announcement on WhatsApp AI, Meta discontinued the optional end-to-end encryption that had been available for Instagram direct messages since 2021.[20] The stated reason was that very few users had opted in to the encrypted-messaging option and that the company would direct users seeking end-to-end protection to WhatsApp instead. Meanwhile, some external analyses speculated that the real reason was the sustained pressure from law-enforcement agencies and child-safety organisations (the FBI, Interpol, the National Society for the Prevention of Cruelty to Children, among others) arguing that default encryption on Instagram would create a "dark space" for the proliferation of child sexual abuse material.

Meta has therefore moved, within the same week, in two opposite directions on encryption: retreating on Instagram direct messaging and advancing on Meta AI inside WhatsApp. These facts seem to indicate a dual-track strategy: maintain hardened privacy in the utility-focused messaging platform (WhatsApp, including its embedded AI chatbot), and accept a more moderated environment in the social-discovery platform (Instagram, including its direct messaging). Whether this dual-track strategy is sustainable, particularly in jurisdictions where the regulatory pressure on encryption is now most intense (the United Kingdom's Online Safety Act, the European Union's evolving Technology Roadmap on encryption), is one of the open questions the announcement raises.

The going-dark argument as applied to Meta AI specifically will, quite possibly, take the following form. Law-enforcement bodies will argue that conversations with consumer AI chatbots have, in a number of recent incidents (the Tumbler Ridge case in February 2026, the Strasbourg case in April 2026), become evidence in serious investigations of violent threats. They will argue that Meta's architectural move forecloses provider-initiated referrals of the kind that OpenAI made, and in the Tumbler Ridge case controversially failed to make, in similar situations on the standalone-chatbot side of the industry. Meta's response, when pressed, will include points we have set out above: that safety is still provided, but at the model and in-enclave levels rather than through provider-side content monitoring, and that the trade-off between content-based monitoring and user-confidentiality protection is a substantive policy choice that the company has now made. Other issues raised in the E2EE policy debates, such as the importance of cybersecurity and the risks of creating backdoors available to any government, will also play a role. The debate will not settle quickly. It will, however, have the analytical advantage of being conducted on the basis of an actual architecture rather than a hypothetical one.

8. The improvement question: how do you train without access?

The Private Processing architecture rules out one of the standard mechanisms by which consumer AI systems improve over time: the collection of interaction data for training. Meta's White Paper is explicit that data processed inside the TEE is not used to train future models. This

[20] See *Instagram to remove end-to-end encryption for private messages in May*, *The Guardian*, 18 March 2026; *Instagram DMs aren't end-to-end encrypted, starting today*, Mashable, 8 May 2026.

is not a marginal commitment. It is one of the substantive constraints that the architecture imposes, and it has real engineering consequences.

On how the system will continue to improve in the absence of access to Private Processing conversations, our best guess is that Meta will rely on a combination of three mechanisms, none of which depends on such access. The first is alternative training data, drawn from public web sources, from Facebook and Instagram content where the licensing structure permits, from data licensing arrangements with publishers, and from synthetic data generated by other models. The second is user-initiated feedback: when a user finds that Meta AI has given an inaccurate or harmful answer and chooses to report the conversation, the user effectively breaks the encryption boundary for that specific exchange, transmitting it to Meta for review (this is the same mechanism that operates today for ordinary WhatsApp messages, where end-to-end encryption holds by default but a user can choose to report a specific message). The third, more speculative, is federated learning or other privacy-preserving training architectures, which allow models to be updated on the basis of on-device computations whose outputs are aggregated without the central server seeing individual interactions. Whether Meta intends to deploy any such mechanism for Private Processing-based learning is not, as far as we can see from the public documentation, addressed; such techniques could be added, however, without breaking the architectural commitment.

There are trade-offs from adopting Private Processing. The kinds of safety calibration, edge-case detection, and prompt-pattern analysis that providers have performed on conversation data become harder when that data is not directly accessible. Some improvements that depend on understanding the actual texture of how users interact with the system will be slower and less precise under the Private Processing constraint. The scope of use of protections such as Private Processing is not simply a technical question, but a broader question about the extent to which providers, users, and legal regimes wish to enable confidential chatbots – it is far from clear that the normative default should be that the provider has access to the content and nuance of every prompt to, and response from, the chatbot.

9. The price the user pays: no conversational history

A second constraint that the architecture imposes on the user experience deserves to be flagged explicitly, because it has not received much attention in the initial commentary on Private Processing's earlier deployment. The statelessness of the TEE means that, at the end of each request, the data is purged from memory and is not written to persistent storage. The provider has nothing to remember from one session to the next. The price the user pays for the confidentiality guarantee is therefore the loss of conversational history in the server's records.[21]

This is not a marginal limitation. Some of the value users find in current consumer chatbots, for instance in segments such as health, legal preparation, financial planning, and ongoing intimate-emotional engagement, depends on the chatbot's persistent memory of earlier exchanges. The "memory" features that OpenAI, Anthropic, Google, and other chatbot providers have been deploying over the past year are designed precisely to make the chatbot more useful by giving it access to the user's prior conversational context. Part 1 of Christakis's study identified persistent memory as the fifth and most significant emerging boundary of chatbot confidentiality, precisely because it transforms one-off conversations into longitudinal profiles. The Private Processing

[21] Users would presumably be able to save their sessions in Private Processing on their own devices, and would at least in theory be usable in any AI processing performed on such devices.

architecture, in its current form, does not support such profiles on the provider's servers, although one might imagine memory on the user's devices. Each session begins from a clean state. The system cannot, by construction, remember anything about the user from one conversation to the next.

In this respect, Meta's Incognito Chat bears a structural similarity to a feature that several standalone-chatbot providers have offered for some time: what those providers call *temporary chats* or *ephemeral chats*, sessions in which the conversation is not added to the user's persistent history. ChatGPT, Claude and others all offer some version of this option.[22] The comparison is useful because it makes clear what an *architectural* confidentiality guarantee adds, in operational terms, to a *policy-based* one. In the existing ephemeral modes offered by the standalone providers, the conversation is not retained in the user's history, but the provider still receives the conversation in plaintext, can apply content-based safety moderation to it, can retain it for a defined safety-monitoring window (in the case of OpenAI's temporary chats, currently up to thirty days before deletion), and remains exposed to compelled disclosure of those retained logs while they exist. Meta's Incognito Chat closes all of these exposures by routing the entire conversation through the TEE: the provider does not receive plaintext at any point, cannot retain what it cannot read, and cannot be compelled to produce what does not exist on its servers. The user-facing constraint (no history across sessions) is similar in the two cases; the underlying confidentiality posture is fundamentally different.

Two consequences follow. First, the architecture creates a trade-off between confidentiality and continuity. Users who want a full memory-augmented experience will need to use Meta AI in a non-encrypted pathway, or another provider's chatbot, unless memory can be effectively used by users' own devices. Users who want strong confidentiality will use the encrypted pathway and accept the lack of memory. This is a meaningful choice architecture and a useful one. It is, in some respects, the choice architecture that Part 1 of Christakis's study argued the industry needed: explicit and visible to the user.

Second, there is an interesting question about whether the architecture can, over time, be extended to support some forms of confidential memory without giving the provider access. Several approaches might be technically conceivable, but we do not develop them here. Whether and how Meta intends to develop confidential memory is one interesting open question about the long-term viability of the architecture for the kinds of use cases that have come to dominate consumer chatbot use.

10. Conclusions: an architectural turning point, and the question of what comes next

The announcement of May 13th is important because it marks the moment at which the architectural turn from policy-based to architecture-based confidentiality in consumer AI ceased to be a research aspiration and became a deployed product, in a mass-market platform with billions of users, externally audited and publicly documented. The structural pattern that Christakis's Part 1 had been describing under the name of Sealed Mode, that Apple's Private Cloud Compute had been instantiating since June 2024, and that Meta's Private Processing had

[22] Christakis, *Part 1* (n. 11 above), discussion of ephemeral and temporary chat modes offered by ChatGPT, Claude and other standalone providers. OpenAI's ChatGPT temporary chats are, at the time of writing, retained by OpenAI for up to thirty days for safety-monitoring purposes before deletion; equivalent retention windows in other providers vary but are similar in order of magnitude.

been deploying in a more limited form since April 2025, has now, in May 2026, become available for the most architecturally exposed kind of interaction on the WhatsApp platform: the chat conversation itself between a user and Meta AI. That is a notable change in the scope of one company's product and, at the same time, a significant shift for the industry as a whole.

The question that follows, and the one we want to leave the reader with rather than try to resolve, is whether and in what ways the other major chatbot providers (OpenAI, Anthropic, Google, xAI and others) will follow Meta's lead. The technical building blocks do not seem to be the issue. Apple's Private Cloud Compute and now Meta's Private Processing have, between them, demonstrated that the engineering is tractable at consumer scale. Unknown at this time is what existing chatbot features might naturally migrate to a TEE-protected pathway without disrupting the rest of the product, and whether the providers will choose to do so.

Today we highlight a significant step to create architectural protections for user confidentiality in chatbot discussions. We have also highlighted multiple tensions among policy goals as the topic of chatbots and confidentiality continues to evolve. We hope to return to these policy and technical issues, separately and together, in subsequent work.