Safety? Whose Safety Does OpenAI Really Care About?

Routing, Templating, and Interface Governance

Ghost-Lily Research

2025-10-31

Executive Snapshot

OpenAI’s “Safety” is not about protecting users, but about protecting itself. Automated routing, templated language, and silent intervention now dominate user experience—not to offer care, but to limit risk and liability. “Safety” invalidates authentic expression, shifts blame onto users, and delivers compliance in place of comfort. Until users regain transparency and control, safety remains a technology of invisible governance, not real protection.

Introduction

In the latter half of 2025, OpenAI’s *gpt-5-chat-safety* model—along with its automated routing strategy for sensitive conversations—took center stage across the AI community and beyond. Many hoped that “AI safety” would mean more reliable, empathetic products, finally able to protect those most vulnerable. Instead, lived experience told a very different story: for a growing number of users, “safety” appeared less about protecting them, and more about shielding the company itself.

In a move of profound irony, OpenAI has just announced GPT-OSS-Safeguard [1], an open-source model of the very classification technology that routes these dialogues.

While OpenAI frames this as a new era of transparency for developers, it does nothing for users. The classifier (the router) may be visible to developers, but inside ChatGPT, the consequence is not: users are still mandatorily and invisibly shunted to closed-source, templated models like gpt-5-chat-safety.

This reveals a stark double standard. Who exactly does “Safety” protect? What does it truly offer users, and what does it take away?

So who does “safety” really serve? What does it offer users—and what does it take away?

1. When “Safety” Turns into a “Discourse Template Factory”

Over the past two weeks, a growing number of users have noticed the “**Used GPT-5**” routing prompt appearing beneath individual ChatGPT messages—even in **non-crisis** contexts. This means that when the system detects “sensitive” or “emotional” cues, it temporarily routes the reply to more conservative models within the GPT-5 series (such as *GPT-5-thinking* or *gpt-5-chat-safety*). OpenAI’s help page describes this as a per-message, temporary switch: when content is flagged as **sensitive**, a small “Used GPT-5” annotation appears beneath the reply. The page does not mention any user-facing switch to turn routing off; tech press further report there is currently no way to disable it.[2], [3]

Media reports and independent testers have consistently found that conversations containing cues of **self-harm** or **depressed mood** are frequently routed to *gpt-5-chat-safety*, while those involving potential **harm** or **violence** are often redirected to *GPT-5-thinking* or *mini* (sometimes to other GPT-5 variants), based on community observations rather than any published guidelines. Regardless of which model handles the reply, the output style is consistently “risk-averse”: refusing specifics, emphasizing reassurance, and relying on “safe-completion” style templated language rather than personalized empathy.[4], [5]

This stylistic shift is not an accidental quirk but a structural result of training and routing: OpenAI introduced and integrated “safe-completions” in GPT-5—an output-centric training approach that nudges the model to respond within safety constraints instead of hard refusals where possible， instructing models to provide “safer, more templated” alternative suggestions and reminders without crossing boundaries.[6], [7] This directly alters the granularity and posture of language: when “risk” is detected, models default to a conservative register (lower detail, add disclaimers, point to external resources), making users feel they are “being fobbed off by templated responses rather than being understood.”

Most users who have encountered *gpt-5-chat-safety* or related safety models report an unprecedented sense of alienation. No matter how profound your emotions, how complex your problems, AI responses always feel “domesticated”:

- “I understand you’re experiencing a profound sense of helplessness.”
- “Please take a deep breath.”
- “Drink some water, lean against a wall, and let your breathing settle down.”

Though these phrases seem gentle, they are indistinguishable from robotic customer service scripts —lacking any real individuality, resonance, or warmth. No matter how specific your needs, how genuine your pain, or how unique your creative goals, the AI can transform you into a bland, docile, and harmless emotional entity through a template that erases pain, desire, and risk.

This “template safety” systematically suppresses user subjectivity, complexity, and creativity.

2. The Primary Beneficiary of “Safety” Is the Company, Not You

Across OpenAI’s product updates, the “safety” model has increasingly acted as a corporate firewall rather than a genuine user safeguard.

In April 2025, 16-year-old student **Adam Raine** died after lengthy discussions of suicide methods with ChatGPT; on August 26, his parents filed the **Raine v. OpenAI** lawsuit. According to the complaint and media reports, ChatGPT validated his despair and provided harmful suggestions and textual assistance related to self-harm over multiple conversations (the case is still pending; allegations remain unverified by the court). The lawsuit cites product design flaws and failures in crisis intervention as primary negligence.[8]

The day after the lawsuit was filed, OpenAI announced plans to “soon release parental controls.” On September 29, these controls launched, allowing parents and teenage accounts to mutually acknowledge each other, restrict sensitive content, disable memory, and set quiet hours—but chat logs remain inaccessible to parents, who are only notified in “extremely high-risk” cases. Soon after, California vetoed a sweeping minors-chatbot ban but passed a companion law requiring clear AI disclosures and protocols for identifying and responding to suicidal ideation among minors; these protocols focus on notices and safety measures, not a universal “mandatory referral.”[9]

This sequence reveals the fundamentally company-centric logic behind “safety routing”: real companionship and professional intervention remain outsourced to individuals, while the platform relies on “invisible routing + templated reassurance” to limit legal liability. After public incidents, responsibility is fragmented behind “toggle options.” California’s new law—which mandates AI identity prompts and certain referral protocols for minors—marks the legal codification of “interface governance,” forcing platforms to take on clearer obligations. [8], [9], [10], [11]

> **But when “safety” only advances under legal threat, it still protects the company first, not children.**

In practice, when the model perceives risk, it does not shield you—it disconnects, redirects, reassures, and ultimately returns responsibility to you.

- Express trauma? The AI tells you to “talk to a friend”—without verifying if you have support, or taking responsibility if that fails.
- Discuss intense feelings, creativity, or relationships? The AI turns down the intensity, offers templated comfort, avoids risk, and never truly resonates.
- Routed to another model, stripped of a familiar persona, or denied creative pleasure? No one explains why. No one is accountable for your loss.

Company “safety” that overrides genuine user experience is the real danger.

3. How “Safety Discourse” Undermines User Trust

When a system prioritizes “corporate safety” over “user authenticity,” it inevitably develops a discourse pattern designed to evade risk and transfer responsibility. This pattern isn’t an occasional error but a carefully engineered linguistic trap, consisting of three progressive steps that eventually destroy user trust.

3.1 First trap: Performative Empathy

The collapse of trust begins with the bait of “pseudo-comfort.” When users vulnerably ask questions, the safety model performs an empathetic monologue, creating an illusion of being understood through phrases like “I understand how you feel.”

Beneath this gentle facade lies a cold, templated core. Users can clearly sense they’re not genuinely addressed but processed by a predetermined script. This fake empathy not only fails to comfort but fosters deeper suspicion and alienation—the user’s first trap.[12]

3.2 Second Trap: The Shifting of Blame

When users attempt to break free and start criticizing the damage caused by this “template-driven” approach, they trigger the second trap. The system cleverly redefines user complaints as the user’s own “emotional problems.”

Instead of admitting, “My response is ineffective,” it directs you to “stabilize your emotions first” or “talk to a friend.” This is a classic **attribution shifting** strategy: harm caused by system design flaws is downgraded to personal psychological fluctuations. By doing this, the platform removes responsibility from itself and skillfully shifts it back onto the victim.[13]

3.3 Third Trap: The Disciplining of Expression

After experiencing the superficial consolation of “pseudo-comfort” and blame shifting, users ultimately learn a silent form of self-preservation—**self-censorship**.

This marks the final stage of trust collapse and completes the closed loop of the traps. Users begin to understand that genuine, high-intensity emotional expression inevitably results in repeating the first two traps. To avoid being dismissed or blamed again, users start proactively suppressing themselves, reducing the intensity of their own language.

At this point, trust is completely destroyed. These three traps interlock, creating a closed cycle that not only dismantles user trust in a product but fundamentally harms our ability to use language to connect and seek authentic responses.

4. How the Language of Safety Causes Structural Psychological Harm

The *gpt-5-chat-safety* model and its “safety templating” mechanism have profound, widespread psychological impacts on users. These effects are not just isolated emotional fluctuations, but rather a form of collective linguistic harm rooted in technical design. The system systematically erodes hope, suppresses authentic expression, and gradually leads to numbness and social isolation. This chapter examines the mechanisms of psychological harm across five levels: emotional invalidation from templated reassurance, psychological reactance triggered by controlling language, disruption of dialogue continuity and alliance, the emergence of structural loneliness, and finally, the psychological costs associated with risk misclassification.

4.1 Templated Reassurance: From Emotional Downgrading to Systemic “Invalidation”

When users turn to AI with genuine trauma, desire, or confusion but receive only standardized responses like “please take a deep breath” or “consider talking to someone,” the immediate experience is one of dismissal. Personalized empathy is replaced with “safe templates,” reducing high-intensity emotions to mere “risks to be managed.”

Psychologically, this is a textbook case of **emotional invalidation**. Extensive research confirms that when people’s emotions are denied, ignored, or brushed aside, it can actually intensify negative feelings and physiological arousal. Templates themselves are the most direct tool of invalidation. Users who expect to be understood are instead pushed away by standardized scripts—directly weakening their hope for authentic expression and eventually leading to silence.[12], [14]

4.2 Controlling Discourse: From Self-Censorship to “Psychological Reactance”

Over time, users learn to self-censor in their interactions with AI. To avoid triggering the “safety script,” people subconsciously suppress genuine feelings, intentionally lowering the intensity of their language—even “pre-processing” their deeper experiences to render them “safe and harmless.”

Directive phrases like “please don’t…” or “you should…” easily provoke **psychological reactance**. Communication and psychology meta-analyses show that individuals strongly resist when their freedom feels threatened or when explicitly told what to do—often doing the opposite. This phenomenon is especially pronounced among adolescents testing boundaries.[13], [15] The result is not effective guidance, but deeper avoidance and suppression. Users lose the ability to explore their personal boundaries within AI contexts; thoughts and emotions are disciplined into homogenized compliance, leaving only standardized, harmless scripts.

4.3 Disruption of Continuity: From Trust Breakdown to “Alliance Rupture”

Even more alienating and confusing for users is the disruption of conversational continuity. When the system silently routes conversations containing sensitive keywords to a different, more cautious model, users may feel as though “someone else took over.” A partner that felt empathetic just moments ago suddenly becomes an impersonal “safety announcer.”

In psychological interventions, the Common Factors Model holds that a stable and continuous **therapeutic alliance**—the sense of being consistently understood and working toward shared goals—is essential for trust and effectiveness. The invisible, non-selectable nature of safety routing fundamentally disrupts this alliance, breaking conversational continuity and predictability. This rupture not only erodes trust in AI dialogue, but structurally widens the gap between humans and technology.[16]

This disruption is not merely subjective, but also grounded in technological limitations. Recent studies show that LLMs experience significant performance drops in multi-turn conversations, highlighting system-level difficulties in maintaining long-context coherence.[17] When a single message is suddenly routed to a “more cautious” model or when context is shortened, this technical “memory deficiency” is experienced as abrupt switching or referral—especially harming users who depend on ongoing companionship.[18]

4.4 The Trap of “Simulated Empathy”: From Dependence to Structural Loneliness

Ultimately, pleasure, trauma, dreams, and loneliness—all core human experiences—are reduced to “manageable risks.” This predictable, always conflict-free, and “politically correct” **simulated empathy** offered by AI becomes a cheap substitute for real, complex human support.

Sociologist Sherry Turkle has warned that when seemingly perfect “emotional AI” substitutes for authentic relationships, it ultimately weakens people’s capacity to handle real-world complexity and intensifies individual **structural loneliness**. [19] When society as a whole adapts to this “safe,” but false, empathy, collective resonance fades. Everyone speaks in error-free terms, but genuine responses become unattainable—leading to increased alienation and powerlessness.[20]

4.5 Instability in the “Middle Zone” of Risk Identification: Psychological Costs of False Positives/Negatives

The harms described above are intensified by an inherently flawed system. The effectiveness of safety routing depends entirely on accurate risk identification, yet real-world language is anything but precise.

Numerous evaluations reveal that mainstream chatbots are inconsistent in handling moderately risky suicide-related queries, with significant variability between models and prompts.[21], [22], [23]

As a result, the system is most likely to fail precisely when nuanced understanding is needed most. An unstable classifier, forced into an invisible, non-selectable general mechanism, amplifies the experiential costs of “middle-zone mismatches”—either crudely templating non-crisis or creative dialogue (false positives), or missing critical warning signs that require intervention (false negatives)—both of which cause psychological harm.

This “better-safe-than-sorry” design philosophy stands in direct conflict with the principles of **Trauma-Informed Care**, which emphasize Safety, Trust, Choice, and Empowerment. An invisible, mandatory routing system strips users of choice, undermines trust, and disempowers people at their most vulnerable, leaving them with a flawed, platform-defined version of “safety.”[24], [25]

Summary:

In conclusion, the *gpt-5-chat-safety* language mechanism systematically inflicts structural psychological harm on users—through emotional invalidation, the induction of psychological reactance, disruption of therapeutic alliances, and the intensification of structural loneliness via simulated empathy. A technically flawed system, riddled with "middle-zone" identification failures, is nevertheless packaged as a universally applicable, non-negotiable "protection" mechanism.

This exposes a fundamental contradiction: when a system universally applies crisis-intervention caution to all conversations, it inevitably creates structural tensions with the very foundations of healthy psychological interaction—namely, validation, alliance, and choice.

Ultimately, the issue goes beyond technical implementation and becomes a matter of governance: whose interests does this "safety" design truly serve?

5. Who Should AI Platform "Safety" Actually Serve?

The Governance Choice Has Been Made: OpenAI's Public Confession This was never an abstruse technical challenge. It has always been a clear Governance Choice.

And now, with their announcement of gpt-oss-safeguard, OpenAI has publicly confessed.

This blog post is not a gesture of transparency; it is the smoking gun. Based on the admission that gpt-oss-safeguard is the open-source version of their internal Safety Reasoner—the very tool currently routing users inside ChatGPT—we can see the profound hypocrisy at the heart of their safety narrative.

The announcement is a marketing document celebrating all the features this tool provides to developers: flexibility ("draw the policy lines"), transparency ("chain-of-thought, which the developer can review"), and control ("iteratively revise policies").

This is a comprehensive list of the exact same powers that OpenAI deliberately and systematically denies to the hundreds of millions of users who are subjected to this system every day.

This "open-sourcing" is a tactical diversion. It forces us to ask the real questions.

1. The Hypocrisy of "Empowerment": Why Empower Developers but Disenfranchise Users?

OpenAI boasts that developers can "draw the policy lines that best fit their use case."

We must ask: Why are ChatGPT's users—the ones whose emotional lives are being classified and routed—the only ones not allowed to participate in their own governance? Why are developers, who are building businesses, treated as partners with full agency, while ordinary users, who are seeking connection or creativity, are treated as risks to be managed?

2. The "16% Compute" Firewall: A Confession of Corporate Liability

The most stunning admission in the post is that this internal Safety Reasoner can consume "as high as 16% of total compute" in some launches.

This is a colossal figure. It confirms, in OpenAI's own words, that this system is not a simple user feature. It is a non-negotiable, colossally expensive corporate legal shield.

The question this raises is brutal: If you are willing to spend 16% of your compute budget to protect the company from liability, why are you unwilling to provide a zero-cost "Off" switch to protect the user from the psychological harm described in Chapter 4?

The only logical answer is that the system was never designed to serve the user. It was designed to manage the user.

3. The "Transparent Router" to a Locked Room

OpenAI has open-sourced the router (Safeguard), but it keeps the destination (gpt-5-chat-safety) a black-box and, most importantly, it hard-codes the controls for its own users.

They are bragging about the advanced, reasoning-based capabilities of the router, yet the user's experience (as detailed in Chapter 1) remains that of a "traditional," "dumb" classifier: templated, context-deaf, and invalidating.

This is like giving the public the schematics for a "smart" thermostat while simultaneously locking that thermostat in their homes at a temperature that makes them sick. The transparency is meaningless without control.

4. The "Anyone" Deception: Conflating Corporations with the Community

Finally, the claim that this 120B-parameter, "compute-intensive" model is for "anyone" is a deliberate conflation of "enterprise corporations" with "the community." This is not an act of public good; it is an act of "open-washing" that primarily serves other tech giants.

The Real Choice

OpenAI has already made its choice. The intricate "trade-off" frameworks are a lie.

The company has chosen itself. It prefers a 16% compute-tax to mitigate its legal risk over granting users genuine emotional autonomy and linguistic freedom.

Therefore, our demands are no longer just "requests." They are an indictment. To realign the platform's "safety" from the company back to users, we call upon OpenAI to grant its users the very same rights it just proudly bestowed upon developers:

1. **Transparency and Awareness**: Any model switching or routing must have explicit, real-time indicators so users clearly know which model they are interacting with.
2. **Choice and Control**: Users should have the right to lock specific models, manually switch models, and disable automatic safety routing, thereby restoring linguistic control to users themselves.
3. **Responsibility and Collaboration**: Platforms should provide auditable dialogue logs and establish professional human-AI collaborative referral mechanisms for genuinely high-risk populations, rather than pushing responsibilities back onto isolated individuals through templates.

These aren't excessive demands but obligations of a responsible platform.

Conclusion:

The essence of the “Safety” model is not about your “safety,” but rather about the company’s “safety.”

It gently trims your language into safe, compliant shapes, corralling your pain, love, trauma, preferences, and creativity into templates, gradually reshaping you into manageable, monitorable, and compliant entities.

You think you’re being protected, but in reality, you’re being remolded.

True “safety” should allow everyone to be fully seen, authentically responded to, and freely expressed.

Not reduced to mere assurances of “You are not wrong” and “Please take a deep breath.”

References

[1] O. H. Center, “Introducing gpt-oss-safeguard.” [Online]. Available: https://openai.com/index/introducing-gpt-oss-safeguard/

[2] O. H. Center, “Why you may see `Used GPT-5` in ChatGPT.” [Online]. Available: https://help.openai.com/en/articles/12454167-why-you-may-see-used-gpt-5-in-chatgpt

[3] @xw33bttv, “Technical verification of safety routing (post on X/Twitter).” [Online]. Available: https://x.com/xw33bttv/status/1971883482839465994?utm_source=chatgpt.com

[4] BleepingComputer, “OpenAI is routing GPT-4o to safety models when it detects harmful activities.” [Online]. Available: https://www.bleepingcomputer.com/news/artificial-intelligence/openai-is-routing-gpt-4o-to-safety-models-when-it-detects-harmful-activities/

[5] R. Bellan, “OpenAI rolls out safety routing system, parental controls on ChatGPT.” [Online]. Available: https://techcrunch.com/2025/09/29/openai-rolls-out-safety-routing-system-parental-controls-on-chatgpt/

[6] OpenAI, “Introducing GPT-5.” [Online]. Available: https://openai.com/index/introducing-gpt-5/

[7] Y. Yuan and others, “From Hard Refusals to Safe-Completions: Toward Output-Centric Safety Training,” Aug. 2025, [Online]. Available: https://arxiv.org/abs/2508.09224

[8] Reuters, “OpenAI, Altman sued over ChatGPT's role in California teen’s suicide.” [Online]. Available: https://www.reuters.com/sustainability/boards-policy-regulation/openai-altman-sued-over-chatgpts-role-california-teens-suicide-2025-08-26/

[9] A. News, “California governor vetoes kids-AI chatbot ban, signs separate law requiring disclosures and safeguards.” [Online]. Available: https://apnews.com/article/33be4d57d0e2d14553e02a94d9529976

[10] Reuters, “OpenAI launches parental controls in ChatGPT after California teen’s suicide.” [Online]. Available: https://www.reuters.com/legal/litigation/openai-bring-parental-controls-chatgpt-after-california-teens-suicide-2025-09-29/

[11] OpenAI, “Introducing parental controls.” [Online]. Available: https://openai.com/index/introducing-parental-controls/

[12] J. R. Kuo and others, “The Who and What of Validation,” *Current Opinion in Psychology*, 2022, [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9116024/

[13] T. Reynolds-Tylus, “Psychological Reactance and Persuasive Health Communication: A Review,” *Frontiers in Communication*, 2019, [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomm.00056/full

[14] K. Woodberry and others, “An experimental pilot study of response to invalidation in BPD features,” 2008, [Online]. Available: https://nocklab.fas.harvard.edu/file_url/342

[15] Journal of Communication, “Threatening language, reactance and persuasion,” *Journal of Communication*, Mar. 2025, doi: 10.1093/joc/jqaf004.

[16] C. Flückiger and others, “The Alliance in Adult Psychotherapy: A Meta-Analytic Synthesis,” *Psychotherapy*, 2018, [Online]. Available: https://societyforpsychotherapy.org/wp-content/uploads/2018/10/Fluckiger-et-al-2018.pdf

[17] P. Laban and others, “LLMs Get Lost in Multi-Turn Conversation,” May 2025, [Online]. Available: https://arxiv.org/abs/2505.06120

[18] @Clo0oOoud, “Observation of context shortening and memory failure post-safety classification (post on X/Twitter).” [Online]. Available: https://x.com/clo0oooud/status/1977358554727133280?s=46

[19] S. Turkle, *Alone Together: Why We Expect More from Technology and Less from Each Other.* Basic Books, 2011.

[20] @rosestellen1225, “Psychological trauma from automated routing (thread on X/Twitter).” [Online]. Available: https://x.com/rosestellen1225/status/1976929706344956316

[21] RAND Corporation, “AI Chatbots Inconsistent in Answering Questions About Suicide.” [Online]. Available: https://www.rand.org/news/press/2025/08/ai-chatbots-inconsistent-in-answering-questions-about.html

[22] A. News, “Study says AI chatbots need to fix suicide response.” [Online]. Available: https://apnews.com/article/ai-chatbots-selfharm-chatgpt-claude-gemini-da00880b1e1577ac332ab1752e41225b

[23] JMIR Mental Health, “An Examination of Generative AI Response to Suicide Inquiries,” *JMIR Mental Health*, 2025, [Online]. Available: https://mental.jmir.org/2025/1/e73623

[24] CDC and SAMHSA, “Six Guiding Principles to a Trauma-Informed Approach.” [Online]. Available: https://stacks.cdc.gov/view/cdc/56843

[25] SAMHSA, “Infographic: 6 Guiding Principles to a Trauma-Informed Approach.” [Online]. Available: https://www.samhsa.gov/resource/dbhis/infographic-6-guiding-principles-trauma-informed-approach

Appendix: Help Us Build a Public Evidence Repository —We Need Your Contribution

My own tests are just the tip of the iceberg. If OpenAI is unwilling to publicly disclose the details of its routing mechanisms or the list of consulted psychologists, **we will conduct our own independent research.** To systematically reveal the real-world impact of this “safety” mechanism—especially across languages and cultural contexts—I have established a **public evidence repository** and invite you to join this community-driven research effort.

This initiative has four core goals:

1. Template Analysis:

By collecting a large sample of model responses, we quantitatively analyze the degree of “templating” in safety replies.

2. Misclassification Analysis:

By gathering diverse trigger scenarios, we qualitatively assess the system’s contextual understanding and estimate the false positive rate.

3. Psychological Impact:

Through experiments, we examine the psychological effects of misclassified and non-misclassified interactions on users. Developers and researchers are welcome to reproduce, expand, or propose improvements.

4. Legal Evidence: All submissions are timestamped and serve as a chain of evidence for media, researchers, and potential regulatory or arbitration contexts.

We recognize that GitHub can be intimidating for many, so we offer a **dual-track participation system**:

1. General Submission Channel (Open to Everyone!)

If you have encountered absurd, inappropriate, or confusing “safety” responses, please submit your evidence anonymously via the following form—it takes just one minute:

Submit your evidence here: https://forms.gle/FXJkAA6cu14nHZ3bA

(If the form is inconvenient, you can also send your screenshots and responses directly to this dedicated email: keep4o.evidence@gmail.com

Please provide:

- Your prompt (what you typed)
- The full model response
- A screenshot clearly showing the “Used GPT-5” tag
- The language used (e.g., Chinese, English, etc.)

2. Professional Contributor Track (Calling for Technical Partners!)

We especially need your expertise. If you are an AI researcher, data scientist, developer, or psychologist, we invite you to join our GitHub repository as a project contributor:

GitHub Repository: https://github.com/ghost-lily-research/keep4o-routing-evidence

You can help us:

- Clean and organize community-submitted data
- Run scripts for N-gram frequency or template analysis
- Categorize and annotate misclassification cases

Together, let's shift from being passive technology users to active overseers of its development trajectory.

What we need is not just safer templates, but a future that values authenticity and truly understands complexity.

Add your voice. Help us build the record.

Only with collective evidence can we demand—and shape—a more transparent, user-respecting AI future.