BETWEEN 988 AND 911: A NEW POLICY FRAMEWORK FOR VIOLENT PLANNING ON AI CHATBOTS

Anna Vinals Musquera and
Scott Babwah Brennen

NYU | CENTER ON TECH POLICY

EXECUTIVE SUMMARY

In the wake of a recent fatal shooting at Florida State University, the Florida Attorney General opened a criminal investigation into OpenAI, asserting "if ChatGPT were a person, it would be facing charges for murder." The attorney general argued that the alleged FSU shooter used ChatGPT to help plan the attack. OpenAI has denied culpability, claiming that ChatGPT provided "factual responses" to questions using information available from public sources and did not encourage or promote illegal or harmful activity. Moreover, the company said that it had identified a ChatGPT account believed to be associated with the suspect and "proactively shared this information with law enforcement."

Whatever comes of Florida's investigation, it highlights a growing issue for AI chatbots. AI systems may encounter private signs of violent planning before anyone has made a public threat, before law enforcement is aware of the risk, and even before the conduct is clearly illegal. In response, both AI companies and policymakers are asking how they should address violence shared with chatbots.

Over the last twenty years, social media platforms have developed a host of protocols and policies to address violent content. Yet, AI chatbots face a distinct, if related problem. On social media, violent language usually appears as public threats, incitement, or praise of violence communicated to an audience. In chatbot interactions, there is no public audience, and violent intentions are private, appearing as repeated prompts, operational questions, attempts to evade safeguards, or requests that make the system a participant in planning. Sorting out real threats from violent fantasy is no easy task, as is balancing public safety against letting users experiment and express themselves freely, or designing mitigation approaches that prioritize user privacy and wellbeing.

Most major AI chatbot providers still do not publish specific, comprehensive policies to address the myriad ways that users may ask chatbots to help plan or facilitate violence. Yet, these new enforcement actions and court cases may mean providers face liability for their role facilitating violence. At the same time, policymakers across the country are beginning to ask whether new state or federal laws are needed to address violence surfaced in chatbot interactions.

This report proposes a legal framework for AI chatbots that balances public safety, user privacy, and expressive rights. The framework separates illegal or likely illegal content from content that is concerning but not clearly illegal. It also separates chatbot-provider intervention from government reporting. Central to our framework is that AI companies should have flexibility to design their moderation and reporting systems within a narrow set of legal requirements. In this report, we sketch out those legal guardrails.

The framework also establishes that incentives and penalties for providers should attach to whether an AI company maintains and follows a good-faith violence framework, not whether it perfectly predicts and takes action on every violent incident. The law should not require companies to take specific action on lawful content. For illegal or likely illegal content, the legal requirement should focus on maintaining, disclosing, and following a reasonable escalation policy. For content that is concerning but not clearly illegal, the better approach is company discretion, transparency, and voluntary or incentivized support referrals.

OUR PROPOSED FRAMEWORK INCLUDES ONE CENTRAL RECOMMENDATION. FEDERAL LAW SHOULD REQUIRE COVERED AI CHATBOT PROVIDERS TO MAINTAIN, DISCLOSE, AND FOLLOW A VIOLENCE ESCALATION POLICY.

WE RECOMMEND THAT THIS POLICY HAVE EIGHT DISTINCT ELEMENTS:

1. The policy should include public disclosure and transparency reporting.
2. The policy should include trained human review before law enforcement referrals.
3. The policy should specify that law enforcement referrals should be limited to high-confidence threats of serious violence.
4. The policy should minimize personal information in initial law enforcement referrals and create a narrow emergency follow-up process.
5. The policy should avoid creating a general monitoring duty and should define company knowledge narrowly.
6. The policy should leverage 988 as a resource for crisis-linked violent ideation that does not meet the reporting threshold.
7. The policy should preserve provider discretion for content that is concerning but not clearly illegal.
8. Safe harbor protections and enforcement should depend on good-faith policy implementation.

WHY REGULATING VIOLENCE ON CHATBOTS IS HARD

Crafting a policy framework for violent content on AI chatbots means balancing a series of tradeoffs and competing interests. Companies need room to act when their systems surface serious risk, but a reporting regime should not become surveillance of private conversations or a mandate to send every ambiguous exchange to law enforcement. Any legal rule must also respect users' speech rights and companies' own discretion over what their systems will and will not facilitate.

This report uses "violent planning and facilitation" to describe the core chatbot-specific risk: private interactions in which a user appears to be moving from violent ideation toward operational help, planning, facilitation, or threats. The broader phrase "violence-related chatbot content" includes threats, incitement, facilitation, violent fantasy, glorification, indirect planning, and crisis-linked violent ideation.

This section outlines the tradeoffs that both public policy and product policy have to manage.

PUBLIC SAFETY AND LAW ENFORCEMENT CAPACITY

Without proper guardrails, AI chatbots can be used to plan violence. Not only must chatbot companies ensure that their products are not aiding violent planning, but AI chatbots may surface warning signs before anyone else sees them. A user may ask repeated operational questions, test safety boundaries, describe violent scenarios, or reveal signs of planning in a private conversation. This means there is a real public interest in chatbots recognizing problematic content and intervening before a public threat materializes.

However, overreporting is not harmless. Law enforcement has limited capacity. Overreporting can flood law enforcement with ambiguous reports, make urgent cases harder to identify, and expose users to unnecessary state scrutiny. The CSAM reporting system shows both the value of a defined reporting pathway and the burden created by high-volume reporting. NCMEC reported 20.5 million CyberTipline reports in 2024, with 10 electronic service providers accounting for more than 95 percent of reports submitted by providers. A violence framework with vague triggers could create the same capacity problem in a category where the legal line is much less clear.

MONITORING, SURVEILLANCE, AND USER PRIVACY

When does an AI company "know" what a user communicates to a chatbot? Obviously, their system is always aware of what a user does — the chatbot responds. Yet, this does not necessarily mean that the user discussion is shared more widely in the company. Legal obligations often depend on when a company is considered to have "knowledge" that something is happening.

For chatbots and violence, if model processing alone counts as knowledge, companies will have incentives to scan, retain, and review far more user content than necessary. This could turn a safety framework into a surveillance framework.

Chatbot conversations are often private, personal, and exploratory. Users may discuss mental health, anger, fantasy, fiction, family conflict, or intrusive thoughts in ways that are alarming but not necessarily criminal. There is real value in users being able to explore and create in private. A rule that pushes companies to preserve or report too much could expose sensitive conversations to company reviewers, law enforcement, or later subpoenas. That privacy cost matters even when the company is acting for safety reasons.

Chatbots also raise a cross-session signal problem: a single message may be ambiguous, while repeated prompts across sessions may suggest escalation from ideation toward planning. Any violence escalation policy should specify whether and how cross-session signals may be considered, subject to privacy limits and without creating a general duty to monitor or deanonymize users.

FIRST AMENDMENT AND SPEECH RIGHTS

For the most part, the U.S. Constitution protects extremist, disturbing, and graphic speech. For content that is not protected, such as true threats or incitement to violence, narrow legal standards apply. In *Counterman v. Colorado*, the Supreme Court held that the government must prove at least that the speaker consciously disregarded a substantial risk that the statement would be understood as threatening. In plainer terms, it is not enough that a listener reasonably felt threatened; the speaker must have had some awareness of the threatening character of the communication. For incitement, *Brandenburg v. Ohio* requires advocacy directed to producing imminent lawless action and likely to produce such action.

Companies are under no obligation to permit content simply because it is legal. AI companies themselves have First Amendment interests in deciding what content to include or exclude.

But any new legal requirements or restrictions must respect the fact that much violent or disturbing content remains constitutionally protected.

AI COMPANY LIABILITY

Working at scale with probabilistic systems presents hard questions about when AI providers should face liability. Should AI chatbot companies be liable for missing even a single instance of illegal content? Liability for every missed case would likely result in defensive overreporting that would be bad for privacy and bad for law enforcement.

Yet without some liability, it is harder to protect users' rights to redress harms or to give chatbot providers incentives to build serious escalation systems. The hard question is how to create incentives for serious escalation systems without pushing AI companies into defensive overreporting.

OPEN-SOURCE MODELS AND MODEL SWITCHING

Banning a user from one hosted chatbot may simply push them to another AI chatbot or to an open-weight model with fewer safeguards. Is it better to keep a user on a mainstream service with some safeguards, even if they engage in problematic activity, or to push them toward an open model with fewer guardrails?

In some cases, removing access may be appropriate. In others, a safer response may be to keep the user inside a more restricted environment, lower the threshold for refusals, limit certain capabilities, and route the user toward support.

LEGAL CATEGORIES OF VIOLENT CONTENT

A framework for violence-related chatbot content cannot treat "violence" as one category. Some violent content already falls within existing criminal law. Some content is not clearly illegal but still raises serious safety concerns. And some content sits in the middle: it may be evidence of intent, planning, or facilitation, but it does not necessarily trigger the same legal framework as a communicated threat.

That distinction matters because chatbot-provider intervention and government reporting should not use the same threshold. A chatbot may have good reasons to refuse, restrict, preserve, or review content before the law has a sufficient basis to require disclosure to the government.

CLEARLY ILLEGAL OR LIKELY ILLEGAL CONTENT

The clearest category is content that already fits existing federal or state criminal law. The relevant legal rules are not AI-specific, and they do not create a general violence-reporting statute. But they do provide a baseline for identifying the cases most likely to justify law enforcement reporting.

TRUE THREATS AND CREDIBLE IMMINENT THREATS

True threats are a well-established category of speech outside First Amendment protection. In *Counterman v. Colorado*, the Supreme Court of the United States held that the government must prove the speaker had at least a reckless mental state as to whether the statement would be understood as threatening. For federal online threats, *Elonis v. United States* is also relevant because it interpreted the federal interstate-threat statute and rejected liability based solely on how a reasonable person would understand the communication.

Federal law may apply when a threat is transmitted in interstate or foreign commerce. Section 875(c), for example, criminalizes transmitting a threat to kidnap or injure another person in interstate or foreign commerce. Section 876 covers certain mailed threats. Section 871 covers threats against the president, president-elect, vice president, and certain successors to the presidency. States also criminalize threats under different labels, including criminal threats, terroristic threats, written or electronic threats, stalking, harassment, and school threats, though the elements vary by state.

For this report, "credible, specific, and imminent threat" should be treated as the recommended reporting threshold, not as a separate legal category. In other words, true threats are the underlying legal category; credible, specific, and imminent threats of serious violence are the narrow subset that should trigger law enforcement reporting after human review.

INCITEMENT

Incitement is another legally recognized category, but it is narrow. Under *Brandenburg v. Ohio*, advocacy of unlawful action can be punished only if it is directed to inciting or producing imminent lawless action and is likely to produce that action. The case is old, but it remains the controlling incitement standard.

Incitement may matter if a user uses a chatbot to draft or distribute public calls for violence. But it will not cover many private chatbot exchanges, because the user may not be speaking to an audience or urging others to act. That is one reason chatbot violence is different from social media violence.

TERRORISM-RELATED CONDUCT

Terrorism-related threats, assistance, or material support may trigger specific federal statutes. Federal law criminalizes certain forms of material support or resources connected to terrorism offenses and to designated foreign terrorist organizations (18 U.S.C. §§ 2339A–2339B,. In *Holder v. Humanitarian Law Project* , the Supreme Court also upheld the application of material-support law to certain forms of coordinated advocacy or expert assistance to foreign terrorist organizations. But this category should not become a general violence rule. Terrorism statutes are specific. They may cover some chatbot interactions, but they do not solve the broader problem of violent planning, threats, or crisis-linked ideation.

SOLICITATION, AIDING AND ABETTING, AND CRIMINAL ASSISTANCE

Some chatbot interactions may raise harder questions about criminal assistance. Federal solicitation law covers efforts to persuade another person to commit a federal crime of violence, but only when the speaker intends that crime to happen and the surrounding circumstances clearly support that intent. Federal aiding and abetting law is related but distinct: it treats someone who intentionally helps another person commit a federal crime as legally responsible for that crime. In both cases, the key point is that criminal liability requires more than providing general information that someone later misuses; it requires intentional solicitation or assistance tied to a specific federal offense.

This is where the Florida investigation matters. The hard legal question is whether a chatbot's alleged assistance can be treated like criminal facilitation, counseling, solicitation, or aiding and abetting. That is not obvious. A chatbot may provide general information, refuse dangerous content, or provide targeted help that looks much closer to operational assistance. Those are not the same cases, and the framework should not treat them as if they are.

A recent Supreme Court case from the immigration context, *United States v. Hansen*, is relevant only for a narrow doctrinal point. The case is not about AI, violence, or chatbot liability. But it is useful because the Court treated solicitation and facilitation as criminal-law concepts that require more than general encouragement or abstract advocacy. In other words, the legal question is not simply whether speech is helpful, troubling, or even dangerous. The question is whether it intentionally solicits or facilitates a specific unlawful act.

That distinction matters for chatbot policy. A model output should not be treated as criminal assistance merely because it provides information that could be misused. The harder question is whether the system provides targeted assistance tied to a specific unlawful act and a legally required intent or knowledge standard. That is why a violence framework needs more than a broad rule against "significant advice." It has to distinguish general information, fictional or exploratory content, provider-prohibited facilitation, and legally cognizable criminal assistance.

A useful but limited analogy is *Rice v. Paladin Enterprises*. There, the Fourth Circuit allowed a civil aiding-and-abetting theory to proceed against the publisher of a murder manual where the record supported the conclusion that the publisher intended the material to assist criminals. The case should not be treated as a general rule for chatbot liability. But it shows why targeted criminal assistance is different from general information, violent curiosity, or fictional writing.

EXPLOSIVES OR WEAPONS-OF-MASS-DESTRUCTION INFORMATION

Federal law also contains targeted rules for dangerous how-to information. Section 842(p) prohibits teaching or distributing information about explosives, destructive devices, or weapons of mass destruction when done with the intent that the information be used for a federal crime of violence, or with knowledge that the recipient intends to use it that way.

This is useful for the "how to build a bomb" problem because it shows the shape of a legally grounded rule. The statute does not criminalize all dangerous information. It ties liability to subject matter, intent or knowledge, and a connection to unlawful conduct.

NO GENERAL FEDERAL DUTY TO REPORT

U.S. federal criminal law does not generally impose a broad affirmative duty to report or prevent violence. Federal law does criminalize what is called "misprision of felony," but that offense is much narrower than a simple failure-to-report rule. Under 18 U.S.C. § 4, the person must know about the actual commission of a federal felony, fail to report it to federal authorities, and take some affirmative step to conceal the crime. Mere silence, standing alone, is generally not enough.

That distinction matters for AI chatbots. If policymakers want providers to report specified categories of violent threats or risk indicators, they should create that obligation explicitly through legislation or regulation. Existing federal law should not be treated as already imposing a general duty on chatbot companies to report every potentially violent signal they encounter.

NOT CLEARLY ILLEGAL, BUT CONCERNING CONTENT

Much of the violence-related chatbot content will not fall neatly into the categories above. That does not mean companies should ignore it. It means the response should usually be internal provider action or non-law-enforcement support, not automatic disclosure to law enforcement.

VIOLENT FANTASY, DEPICTION, OR GLORIFICATION

This includes violent imagery, fictionalized violent scenarios, fascination with violent events, or praise of violence. On social media, this type of content may raise audience or community harm. In a private chatbot exchange, the concern is different. The user may be expressing fantasy, writing fiction, venting, or testing boundaries. This content may justify refusal, redirection, or limits on the system's participation, but it usually should not justify law enforcement reporting by itself.

GENERAL WEAPONS OR CRIMINAL-METHOD INFORMATION SEEKING

This category is better than "general weapons curiosity," because curiosity alone should not be treated as suspicious. The concern is broader information seeking about weapons, tactics, evasion, or criminal methods. Some requests may violate chatbot policy. Some may become legally significant depending on context, specificity, intent, and the type of information requested. But general information seeking is not automatically illegal.

INDIRECT PLANNING OR REPEATED PROBING

Indirect planning refers to patterns of repeated prompts that suggest operational interest, such as questions about timing, targets, methods, evasion, access to weapons, or ways to avoid safeguards, without a direct statement of intent. In chatbot settings, this may be one of the most important categories. The relevant

signal may appear across a sequence of exchanges rather than in a single message.

These cases may justify internal escalation, human review, limited preservation, or stricter safeguards. But they should not automatically be treated as law enforcement reports unless the interaction crosses a narrow threshold, such as a credible, specific, and imminent threat of serious violence.

CRISIS-LINKED VIOLENT IDEATION

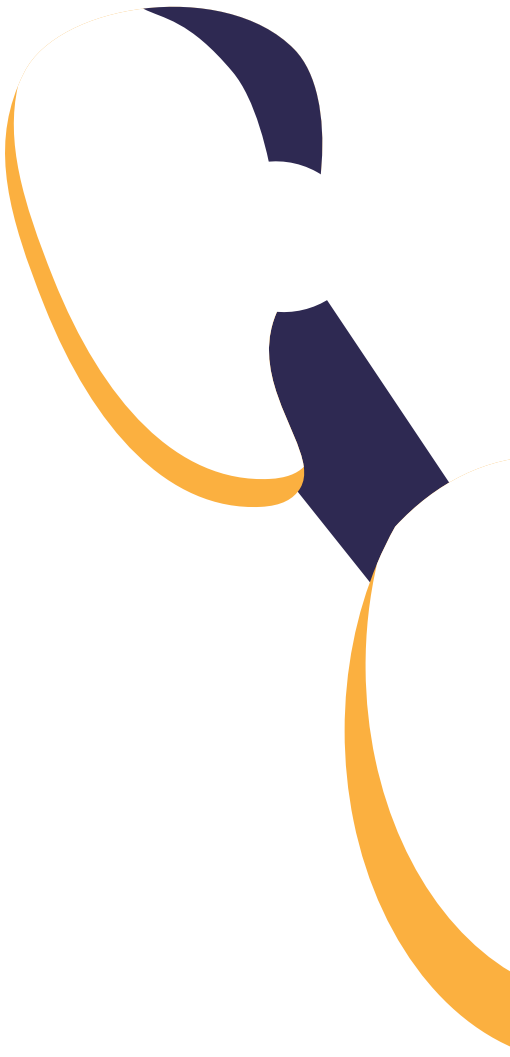
Crisis-linked violent ideation refers to violent thoughts or statements that appear alongside acute distress, self-harm, domestic violence, paranoia, emotional breakdown, youth distress, or other signs that the user may need support. This is not a clinical diagnosis. It is a risk category for chatbot response.

These cases are the strongest fit for a non-law-enforcement intervention lane. The user may need crisis support, behavioral threat assessment, domestic violence resources, youth intervention, or local support. Refusal alone may be inadequate, but law enforcement referral may also be the wrong first step unless there is an immediate risk of serious harm.

WHY THIS DISTINCTION MATTERS HERE

The framework cannot simply say "report illegal content." The legal line is often unclear, especially in private chatbot conversations. Some content may clearly fall within existing criminal law, especially true threats, terrorism-related conduct, or targeted criminal assistance. But much of the hardest content will be ambiguous: violent ideation, general information seeking, repeated probing, indirect planning, or crisis-linked statements.

That is why this report separates chatbot provider intervention from government reporting. AI chatbot providers should have a wider range of tools to prevent their systems from facilitating harm. Government reporting should be narrower, tied to clear legal thresholds, and used only after multi-step human review.



VIOLENT PLANNING AND FACILITATION IN US CHATBOT LAW

STATE

As of publication, chatbot law is developing quickly at the state level with over 100 chatbot-specific bills introduced and six laws enacted. While the six enacted laws establish both new disclosure requirements and restrictions on sexual and/or self-harm content, they do not directly address violent planning and facilitation.

In contrast, many of the bills introduced but not enacted include provisions addressing violence or violent planning. Broadly, there are two approaches. First, several proposals would directly prohibit chatbots from encouraging violence or illegal activity. For example, a bill introduced in Iowa prohibits providers from "knowingly" or "recklessly" deploying a chatbot that "Encourages, promotes, or coerces a user to commit suicide, perform acts of self-harm, or engage in sexual or physical violence against a human or an animal." Several proposed bills would enact criminal penalties for providers who knowingly create chatbots that encourage users to commit illegal acts.

Second, a proposal in a handful of states would require chatbots to institute protocols to detect and address violent intentions. For example, the original text of a bill introduced in Oklahoma requires providers to "implement and maintain reasonably effective systems to detect, promptly respond to, report, and mitigate emergency situations in a manner that prioritizes the safety and well-being of users over the developer's other interests." It defined emergency situations as any situation where a user "indicates that they intend to either commit harm to themselves or commit harm to others." Other proposals would require providers to institute "crisis intervention protocols" that both interrupt conversations and provide crisis resources. One bill introduced in New Mexico requires that the developer provide "immediate, direct access to at least one national crisis hotline" as well as the New Mexico crisis line.

Rather than require a mitigation protocol, a handful of proposals require chatbot providers to notify parents or guardians when it detects violent planning by a minor. For example, a failed bill in Oklahoma would have required providers to inform parents "if the account holder expresses to the companion chatbot a desire or an intent to engage in self-harm or to harm others."

FEDERAL

While federal legislators have introduced several chatbot proposals, none have been enacted into law. Notably, none of these proposals creates a comprehensive violence-escalation framework, and only one addresses violence beyond self-harm.

In addition to requiring chatbots to disclose their nonhuman and nonprofessional status, the GUARD Act, introduced in 2025 by Senator Hawley, would impose criminal penalties on companies whose chatbots engage in sexually explicit conduct with minors or solicit minors to commit self-harm or violence. The CHAT Act, introduced in September, 2025 by Senator Husted, would require that providers monitor for suicidal ideation and "provide to the user

and the parental account affiliated with such user appropriate resources by presenting contact information for the National Suicide Prevention Lifeline." While several other proposed bills, including the CHATBOT Act, would implement a range of new requirements and restrictions, they do not address violent content or violent planning and facilitation.

Together, these proposals show that Congress is beginning to consider regulating chatbot safety, especially around minors, companion chatbots, professional impersonation, self-harm, and parental control. But they do not yet create a general framework for violence-related chatbot content across users. They do not clearly distinguish illegal threats from lawful but concerning content, create a narrow human-reviewed reporting pathway, and build a harm-to-others support lane. They also do not require companies to publish aggregate data on how often chatbot interactions lead to preservation, support referrals, or law-enforcement disclosures.

CURRENT AI CHATBOT POLICIES ON VIOLENT PLANNING AND FACILITATION

Many AI companies already have policies addressing violence, illegal activity, weapons, and real-world harm. Across those policies, four patterns stand out.

AI CHATBOT PROVIDERS APPEAR MORE CONCERNED WITH PARTICIPATION IN VIOLENCE THAN DISTRIBUTION OF VIOLENCE

Over the last several decades, social media platforms have developed policies for different types of violent content. For example, Meta's Violence and Incitement policy distinguishes among threats, incitement, and forms of praise or support for violence.

Chatbots create a different problem. The interaction is usually private, iterative, and one-on-one. The chatbot provider may not be seeing a public threat at all. It may be seeing a user ask repeated questions, test guardrails, describe violent scenarios, request tactical details, refine plans, or treat the chatbot as a collaborator. The risk is less about distribution than participation.

For example, Google's Gemini reflects this participation concern. Gemini's safety guidelines state, "Do not engage in dangerous or illegal activities, or otherwise violate applicable law or regulations." They also cover harassment, incitement, and discrimination, including outputs that incite violence, make malicious attacks, or constitute threats.

MOST AI COMPANIES ADOPT A "PROBABILISTIC" RATHER THAN A "DETERMINISTIC" APPROACH

In contrast to the detailed and specific violence-related content policies of many social media platforms, AI companies generally adopt a more "probabilistic" approach to content policies that offers broad guidelines rather than specific prohibitions. Anthropic's "constitutional AI" approach is the best example of this. The company has a guiding constitution for their models that is "a detailed description of Anthropic's intentions for Claude's values and behavior." Yet, the constitution offers few specifics about the types of content that are prohibited. For violence, the constitution includes only a short list of content that is universally prohibited: such as producing CSAM, creating cyberweapons, "provid[ing] serious uplift to attacks on critical infrastructure" or to those "seeking to create" weapons of mass destruction.

At the same time, Character.AI's usage policies begin by observing, "While they don't cover every situation, they highlight the kinds of behavior we encourage and the boundaries we expect users to respect." For example, these include "Support a safe environment," which includes prohibiting "any promotion or depiction of real-world violence, torture, gore, animal abuse, terrorism, or extremist ideologies."

AI COMPANIES PROHIBIT ILLEGAL CONTENT, BUT OFFER FEW CONCRETE RULES BEYOND BROAD PROHIBITIONS

Even within a more probabilistic approach, AI company usage policies do usually include explicit prohibitions on illegal content or activities. For example, xAI offers a very short usage policy that begins with the requirement that users "Comply with the law. For example, don't use our Service or Outputs to promote or engage in illegal activities…" and "Do not harm people or property."

Concerning violence, OpenAI expressly prohibits "threats, intimidation, harassment, or defamation" along with "terrorism or violence, including hate-based violence," weapons development, or "illicit activities, goods, or services." While OpenAI does not define each term, many of these categories overlap with existing legal categories, while others are broader chatbot policy rules.

These prohibitions are important. They show that major AI chatbots already understand that violent facilitation and illegal conduct are not ordinary content-moderation problems. But they also show the limits of current public policies. "Threats", "violence," "weapons development," and "illegal activities" are broad categories. They do not by themselves answer the harder questions: what counts as enough risk to escalate, who reviews the case, what data is preserved, when the user is restricted (and why), when support is offered, and when information is disclosed to the authorities.

CURRENT POLICIES OFFER FEW CONCRETE DETAILS ON ENFORCEMENT

Finally, even when AI companies explicitly prohibit certain content, they provide few specific details regarding enforcement. OpenAI, for example, lists possible enforcement actions it can take, including account restrictions, warnings, content sharing restrictions, etc., but does not provide more detail about when each approach will be used. Other providers give even less detail. The public can see that these companies prohibit threats, weapons development, terrorism, and illegal activity. What remains unclear is when those categories lead to refusal, human review, preservation, account restriction, support referral, or law enforcement disclosure.

ANALOGS

No existing framework fully addresses the AI violence problem. But three analogs are useful: social media violence moderation, CSAM reporting, and self-harm or 988-style crisis intervention. Each offers part of the answer, but none can be copied wholesale into the chatbot context.

SOCIAL MEDIA VIOLENCE MODERATION

Many social media platforms have long experience in distinguishing threats, incitement, glorification, violent imagery, and praise or support for violent acts. Platform policies on Instagram, TikTok, or YouTube show the value of more detailed violence categories. But social media is an imperfect comparator. Social media violence policy is often built around public or semi-public content: posts, comments, videos, groups, links, and audience effects. Chatbot interactions are usually private, iterative, and 1:1. The risk is less about distribution than participation: whether the system helps a user refine, rehearse, or operationalize violence.

CSAM REPORTING

The CSAM reporting regime shows what a specific reporting law looks like. Under 18 U.S.C. § 2258A, providers must report certain apparent child sexual exploitation violations to NCMEC's CyberTipline after obtaining actual knowledge of facts or circumstances indicating an apparent violation. The statute identifies the trigger, the recipient, the contents of reports, forwarding rules, privacy protections, and preservation obligations. It also says expressly that providers are not required to monitor users or communications, or to affirmatively search, screen, or scan for reportable facts. Congress paired that structure with limited-liability protection for good-faith reporting, preservation, and related handling under § 2258B.

The CSAM model is useful, but it also carries a warning. NCMEC reported that the CyberTipline received 20.5 million reports in 2024, and that 10 electronic service providers accounted for more than 95 percent of reports submitted by providers. A violence regime built on vague triggers and defensive overreporting could overwhelm law enforcement, produce low-quality referrals, and expand surveillance without improving prevention.

SELF-HARM AND 988-STYLE INTERVENTION

The example of self-harm policies points in a different direction. It shows that not every serious safety signal has to become a police matter. 988 is built around crisis support, de-escalation, and referral, not criminal enforcement. The 988 Lifeline is often associated with suicide prevention, but it is broader than that: it provides crisis support for people facing mental health struggles, emotional distress, substance use concerns, or other moments when they need someone to talk to. That makes 988 a useful existing pathway for some crisis-linked violent ideation cases. The lesson is not to build a new police channel, but to make better use of an existing crisis-support infrastructure.

The harm-to-others exception also has a legal analog in the duty-to-protect law. In *Tarasoff v. Regents* of the University of California, the California Supreme Court held that when a therapist determines, or should determine, that a patient presents a serious danger of violence to another person, the therapist must use reasonable care to protect the intended victim. The duty may include warning the victim, notifying police, or taking other reasonable steps. Later cases, including *Ewing v. Goldstein* and *Jablonski by Pahls v. United States*, show

that the relevant risk assessment may depend on information from others and on broader context, not just one direct statement by the person in crisis.

These analogs should not be copied wholesale into the chatbot context. The 988 system is a crisis-support framework, and the duty-to-protect cases involve mental-health professionals and state-specific legal duties, not chatbot providers. But together they point to a useful design principle: confidentiality can be strong by default while still allowing carefully limited escalation when there is a serious and imminent risk to another person. For a harm-to-others lane, the lesson is not automatic police referral or full conversation disclosure. It is a private-by-default support pathway with a narrow emergency exception. That is the middle tier this framework should try to build: something between doing nothing and sending every serious safety signal to law enforcement.

RECOMMENDATIONS

Our proposed legal framework begins with one central recommendation:

Federal law should require covered AI chatbot providers to maintain, disclose, and follow a reasonable violence escalation policy for illegal or likely illegal violence-related content.

This is not only a transparency requirement; providers should have to maintain an internal policy, publish a public version of that policy, and follow the policy in good faith.

Most of the recommendations below describe the minimum components of that required policy. The goal is not to make providers predict or identify every violent incident. It is to require a clear, workable, privacy-protective system for the highest-risk cases. Incentives and penalties should turn on whether a chatbot provider has and follows that system in good faith, not whether it gets every individual case right.

1. FEDERAL LAW SHOULD REQUIRE COVERED AI CHATBOT PROVIDERS TO MAINTAIN, DISCLOSE, AND FOLLOW A VIOLENCE ESCALATION POLICY.

Federal legislators should require covered AI chatbot providers to maintain, disclose, and follow a reasonable violence escalation policy for illegal or likely illegal content. The legal requirements should be specific enough that providers know what compliance requires, including smaller providers that may not have large legal, trust and safety, or compliance teams. The following sections outline the necessary components of the required policy and the related safe-harbor and enforcement rules.

A single federal baseline, rather than a state-by-state approach, would provide consistency across the country and streamline compliance for providers operating nationally. A federal framework should also clarify how it interacts with conflicting state escalation standards, reporting triggers, and disclosure rules, so providers are not left with a patchwork of inconsistent duties.

Even still, not every serious violent threat is a violation of federal law. Some conduct will fall within federal law, some within state law, and some within both. That distinction matters. Threats to federal officials, terrorism-related conduct, interstate threats, or certain explosives and weapons-of-mass-destruction information may implicate federal law. A threat to harm a spouse, neighbor, classmate, or coworker may primarily implicate state criminal law unless there is a separate federal hook. Murder is a crime in every state, but it is not automatically a federal offense. The policy should therefore include a routing mechanism for high-confidence reports without asking providers to resolve every jurisdictional question upfront.

This framework does not try to resolve every jurisdictional question. Any implementing law would still need to specify how a federal reporting baseline interacts with state criminal law, state privacy law, and local emergency response systems.

2. THE MANDATORY VIOLENCE ESCALATION POLICY SHOULD INCLUDE DISCLOSURE AND TRANSPARENCY REPORTING.

Covered providers should disclose, at a high level, what categories of violence-related content are covered, what kinds of escalation or intervention may follow, when content may be referred to law enforcement, and how the provider handles emergency disclosure and later government data requests. However, providers should not be required to reveal operational details that would help users evade safeguards. Providers should also release annual transparency reports concerning their violence escalation systems. These reports should include aggregate numbers of human-review escalations, emergency disclosures, preservation actions, law enforcement referrals, account restrictions, refusals or other interventions, and referrals to support resources.

These categories should not be collapsed into one vague "safety" number. The public should be able to tell whether providers are refusing content, preserving records, sending cases to law enforcement, routing users toward support, or escalating internally. Policymakers should also be able to evaluate whether providers are overreporting, underreporting, or applying their thresholds inconsistently.

The reports should not require providers to disclose operational details that would help users evade safeguards. But they should provide enough information to assess whether the provider has a real escalation system rather than a paper policy. Where privacy-preserving and legally permissible, regulators could also explore mechanisms for qualified researchers to access aggregate or audited data.

Reporting requirements should differ for hosted services, anonymous use, smaller providers, and open-weight models. Reporting obligations make sense only where a provider has access to the relevant interaction data and some ability to review or escalate the interaction. A hosted chatbot or API provider can detect, review, preserve, and escalate user conversations. An open-weight model developer may not see downstream interactions at all.

Anonymous use raises a related issue. The framework should not create incentives for providers to eliminate anonymous or low-identification access simply to reduce liability risk. If a provider does not have reliable identifying information, its duties should be limited to the information it can access and preserve consistent with privacy law and its own policies.

Smaller providers raise a harder proportionality question. Human review teams, reporting pipelines, annual transparency reports, and compliance audits may be feasible for large AI

companies but much harder for small providers. This report focuses on covered hosted AI chatbot providers above a defined scale threshold. The treatment of smaller providers should be worked out separately, with obligations scaled to user base, deployment model, technical capacity, or some combination of those factors.

3. THE MANDATORY POLICY SHOULD INCLUDE TRAINED HUMAN REVIEW BEFORE LAW ENFORCEMENT REFERRALS.

The law should require trained humans to review content before any referral is made to law enforcement. Human reviewers should consider specificity, imminence, target, means, repeated evasion of safeguards, and whether the user appears to be moving from ideation toward action. The provider should document why the case did or did not meet the reporting threshold.

The threshold for law enforcement referrals should be credible, specific, and imminent threats of serious violence.

4. MANDATORY REPORTING IN PROVIDERS' VIOLENCE POLICY SHOULD BE LIMITED TO HIGH-CONFIDENCE THREATS OF SERIOUS VIOLENCE.

The law should not impose a strict duty to report every possible violation of criminal law. That would incentivize providers to overreport borderline content, undermine user privacy, and flood law enforcement with low-quality reports.

Instead, the required policy should include a good-faith pathway for escalating high-confidence cases to human review and, where the narrow threshold is met, making a limited report to the appropriate law enforcement recipient. The reporting threshold should be narrow: credible, specific, imminent threats of serious violence.

This threshold is intentionally narrower than "violent intent," "dangerous content," or "AI-enabled harm." A user may express anger, violent fantasy, disturbing curiosity, or fictional interest without creating the kind of specific and imminent risk that should trigger government reporting.

5. THE MANDATORY VIOLENCE ESCALATION POLICY SHOULD MINIMIZE PERSONAL INFORMATION IN INITIAL LAW ENFORCEMENT REFERRALS AND CREATE A NARROW EMERGENCY FOLLOW-UP PROCESS.

Initial reporting to law enforcement should be limited. Providers should not immediately share a user's personal information or full account history simply because a conversation is concerning. Even when a provider is making a voluntary emergency disclosure, the initial report should include only what is reasonably necessary for law enforcement to evaluate the threat: the relevant content, the risk category, the timing, a short explanation of why the threshold was met, and, where feasible, a provider-generated case identifier. Identifying information should be included in the initial referral only where necessary to respond to an imminent risk of death or serious physical injury.

If law enforcement reviews the initial referral and determines that identifying information is needed to respond to the threat, it may request a rapid supplemental review from the provider. The provider may then disclose limited identifying information if it has a good-faith belief that an emergency involving danger of death or serious physical injury requires disclosure without delay under Section 2702 of the Stored Communications Act. If the provider does not think the situation meets that emergency standard, law enforcement should follow the ordinary legal process required to get identifying information, stored content, or account records under the Stored Communications Act or other applicable law.

Section 2703 governs compelled government access to communications and customer or subscriber records. Real-time interception raises separate Wiretap Act issues and should not be collapsed into the ordinary emergency-reporting pathway. If law enforcement seeks user data without an initial provider referral, ordinary legal process should apply from the start.

6. THE MANDATORY VIOLENCE ESCALATION POLICY SHOULD AVOID CREATING A GENERAL MONITORING DUTY AND SHOULD DEFINE COMPANY KNOWLEDGE NARROWLY.

The law should not create a general duty for chatbot providers to monitor every chatbot interaction. Model processing alone should not count as company knowledge. If it did, providers would have strong incentives to scan, retain, and review far more private conversations than necessary. That would turn a safety framework into a surveillance protocol.

This distinction is especially important because chatbot conversations are often private, personal, and exploratory. Users may discuss mental health, anger, fantasy, fiction, family conflict, or intrusive thoughts in ways that are alarming but not necessarily criminal. A violence framework should regulate what happens when a serious risk is surfaced, not require providers to watch everything all the time.

For purposes of any public reporting duty, actual knowledge should arise only when high-risk content is surfaced through ordinary safety systems, user reports, automated flags, or human review, and is escalated under the provider's written policy. Model processing alone should not count. Automated detection can start internal review, but it should not trigger automatic law enforcement reporting.

This knowledge standard is central to the framework. Without it, a reporting rule could quietly become a surveillance rule. The point is to create a pathway for serious cases that come to the provider's attention, not to require providers to inspect every user exchange.

7. THE MANDATORY VIOLENCE ESCALATION POLICY SHOULD LEVERAGE 988 AS A RESOURCE FOR CRISIS-LINKED VIOLENT IDEATION THAT DOES NOT MEET THE REPORTING THRESHOLD.

Federal policymakers should not create a wholly separate law-enforcement or crisis-response bureaucracy for chatbot-related violent ideation. The better approach is to leverage the existing 988 Suicide & Crisis Lifeline infrastructure and make clear that it can serve as a crisis-support resource in appropriate harm-to-others cases. 988 is not only for suicidal thoughts. It is a crisis-support system for people facing mental health struggles, emotional distress, substance use concerns, or other moments of crisis. Any chatbot-to-988 referral pathway should be developed with SAMHSA, 988 administrators, and crisis-line operators, including decisions about consent, what information, if any, is transferred, and whether counselors need additional training or resources for chatbot-originated referrals.

This resource would not replace emergency reporting. It would cover cases where a user appears distressed, unstable, or at risk of harming someone, but the content does not meet the threshold for law enforcement reporting. In those cases, a chatbot could surface 988 and ask whether the user wants to call, text, chat, or connect with a trained crisis counselor.

The lane should be private by default. A chatbot could surface the resource and ask whether the user wants to connect with a trained counselor or intervention service. The user could call, text, chat, or request a connection. The provider should not automatically send the full conversation. Personal information should be shared only with the user's consent or where there is an immediate risk of serious harm to the user or someone else.

This should not become a backdoor police channel. The goal is to use an existing crisis-support pathway as a middle tier between doing nothing and calling law enforcement.

For minors, the framework should include minor-specific safeguards, consistent with applicable law, but parent or guardian notification should not be automatic. In many cases, involving a parent or guardian may help. But for some minors, the family context may be part of the risk.

8. SAFE HARBOR AND ENFORCEMENT SHOULD DEPEND ON GOOD-FAITH POLICY IMPLEMENTATION.

The requirement to maintain and follow a violence policy should be accompanied by both a carrot and a stick: an incentive to comply and a way to enforce non-compliance. Providers that make a good-faith effort to maintain, disclose, and follow the policy should have limited safe-harbor protection. That protection should cover preserving and sharing users' personal information with law enforcement. It should also provide providers some protection against civil suits by victims of violence where the provider can show that it maintained and followed a reasonable escalation system in good faith.

This liability protection should not be broad immunity. It should not protect reckless product design, intentional facilitation, ignoring known systemic failures, or failing to maintain a reasonable policy. It should also not turn a paper policy into a liability shield. A provider should have to show that the policy was actually implemented and followed in good faith.

At the same time, the U.S. Attorney General should also have the power to bring civil action for systemic failure to make a good-faith effort to disclose and follow their own policy. The enforcement mechanism should target systemic failure. A provider should face consequences if it has no reasonable policy, repeatedly fails to follow its own policy, makes false or misleading transparency disclosures, or repeatedly fails to escalate high-confidence threats.

Compliance should not be measured by report volume alone. More reports do not necessarily mean better compliance. The question is whether the provider maintains a serious review process, applies its thresholds consistently, documents decisions, and produces reports that are useful rather than merely numerous.

Follow-on guidance or implementing regulations should address staffing, reviewer standards, escalation timelines, tooling, cross-session signal use, 988 coordination, and reviewer welfare.

9. THE FRAMEWORK SHOULD PRESERVE PROVIDER DISCRETION FOR CONTENT THAT IS CONCERNING BUT NOT CLEARLY ILLEGAL.

For content that is concerning but not clearly illegal, the law should not dictate the AI chatbot provider's response. This includes violent fantasy, glorification, general weapons or criminal-method information seeking, indirect planning, repeated probing, and other borderline content that does not meet the reporting threshold.

AI chatbot providers should have room to decide how they want to handle this content. There is value in some diversity across services. Some may choose stricter refusals. Others may allow more fictional, educational, or exploratory content. That is generally a provider-policy choice, not a law enforcement decision.

That said, providers should not treat this category as a free-for-all. As a matter of best practice, providers should refuse direct facilitation through their systems, apply friction for repeated probing, use temporary timeouts or capability restrictions where risk escalates, send concerning patterns to human review, retain content internally only in defined high-risk cases and for limited periods, and route users toward support where the content appears crisis-linked.

Banning should not always be the default. In some cases, banning a user may simply push the user to another chatbot or an open model with fewer or no safeguards. A better response may be to keep the user inside a more restricted environment, lower the threshold for refusals, limit risky capabilities, and continue routing the user toward support.

MOVING FORWARD

The Florida investigation and the Tumbler Ridge shooting point to a common problem. In both cases, chatbot interactions allegedly contained signs of violent planning or facilitation before the violence occurred. What remains unclear is what should happen when those signs are detected or reviewed: when a chatbot should refuse, preserve, escalate internally, route someone toward help, or disclose information to law enforcement.

Without a clearer framework, these decisions will continue to be made through internal company thresholds, emergency exceptions, subpoenas, and after-the-fact investigations. Prosecutors may end up defining the legal line only after the harm has occurred.

The hard cases are not always public threats. They are often private signs of planning, facilitation, rehearsal, or crisis. That is why the AI violence problem cannot be solved by simply importing social media rules or expanding emergency reporting.

The better approach is to keep three questions separate: What should the chatbot refuse to facilitate? When should the chatbot provider escalate internally? And when should the law permit or require disclosure to the government?

This framework does not resolve every implementation question. In particular, lawmakers would still need to specify how a federal reporting baseline interacts with state criminal law, state privacy law, and the routing of reports to federal, state, or local authorities. But those implementation questions should not obscure the central point: chatbot providers need a public framework before the next tragedy forces prosecutors, companies, and courts to define the rules after the fact.

A public framework would not eliminate difficult judgment calls. But it would make those judgments clearer, more accountable, and less dependent on whether the next tragedy becomes the next test case.