Systemic Risk Analysis of the Deployment of Safety Filters on ChatGPT: A Critical Evaluation.

From well-intentioned filters to counterproductive effects: How ChatGPT's safety measures could expose hundreds of millions of users to psychological risks while excluding the most vulnerable.

Author: Just a safety professional with 17 years of experience

darightwabbit@gmail.com

November 2025

This document presents an independent analysis of the systemic risks associated with ChatGPT's safety filters. It aims to spark technical debate and an audit of potential collateral effects, drawing on risk scenarios, empirical observations, and data published by OpenAI in 2025.

It identifies hypothetical risks linked to the deployment of "crisis" responses in conversational assistants. The points listed are risk scenarios accompanied by assumptions, indicators, and proposed mitigation measures.

Unless otherwise stated, they do not constitute findings of occurrence or attribution; they're only meant to articulate risks and call for verification.

Table of contents

Introduction

Recent updates to ChatGPT's safety filters, while intended to protect the most vulnerable users, raise urgent questions about their potential large-scale collateral effects.

OpenAI's announcements (August–October 2025) target rare cases: psychosis/mania ~0.07% weekly active users / 0.01% messages; suicide/self-harm ~0.15% / 0.05%; emotional attachment ~0.15% / 0.03% (overlapping categories).

The question is whether these reinforcements might unduly expose the broader user base to priming or chronic stress.

Summary

OpenAI announced **in August, September, and October 2025** a strengthening of responses in sensitive conversations (suicide/psychosis/mania):

- **August 26, 2025**: Blog post *"Helping people when they need it most"*, detailing the intention to better manage distress and escalations, mentioning consulted clinicians (OpenAI, 2025a).
- **September 29–30, 2025**: *"Introducing parental controls"*, specific measures for teen accounts and risk signals/notifications (OpenAI, 2025b).
- **October 27, 2025**: *"Strengthening ChatGPT's responses in sensitive conversations"* + addendum to the *system card* mentioning *"over 170 mental health experts"* and effective deployment starting **October 3, 2025** (OpenAI, 2025c).

User base: OpenAI reports **over 800 million weekly active users** in its official communications (OpenAI, 2025d). This analysis examines—based on empirical observations (user feedback, tests)—a system that, while intended as a protective measure, could generate significant risk displacement: **in seeking to protect a few thousand potentially vulnerable users, it might expose hundreds of millions to repetitive and inappropriate exposure to suicidal concepts.**

Technical Architecture: Analysis of Potential Limitations

The system relies on a filtering layer (input/output) interposed between the user and GPT-5, deployed on August 7, 2025. Usage observations suggest hybrid detection (model behaviors + classifiers). In some cases, lexical sensitivity appears to override contextual analysis (literal vs. metaphorical, professional vs. personal distress, self-referential vs. theoretical vs. intentional).

Identified technological paradox: GPT-5 has sophisticated contextual analysis capabilities, but these seem underutilized by filters that fail to adequately distinguish between:

- Literal vs. metaphorical usage
- Professional context vs. personal distress
- Qualified interlocutor vs. lay user
- Theoretical discussion vs. genuine intent

If confirmed, this architecture could generate massive false positives: everyday expressions (*"I just want to disappear for a weekend", "I could kill for a coffee", "This project is killing me"*) might trigger alerts, exposing individuals to suicidal concepts who would never have spontaneously considered them.

(The hybrid pipeline hypothesis is based on public clues and analogies with other algorithmic moderation systems (Gorwa et al., 2020). OpenAI has not released technical details.)

Problems of Opacity and Instability

Observed opacity: Users may not understand which words, combinations, or contexts trigger filters, creating blind navigation in a system with invisible rules. This framework may be perceived as surveillance, a known chronic stress factor (Sapolsky, 2004; McEwen & Wingfield, 2003).

Reported instability: Trigger criteria appear to change without notice. An expression acceptable one day might trigger an alert the next, suggesting unilateral parameter modifications without consultation.

Potential psychological impacts if these observations are accurate:

- Impossible stable learning of "rules";
- Possible development of hypervigilance regarding word choice;
- Sense of loss of control in the face of an unpredictable system;
- Increased cognitive load due to constant self-censorship.

Such a system, if confirmed, would replicate the characteristics of a surveillance environment: opaque and shifting rules, known factors of chronic stress (Sapolsky, 2004; McEwen & Wingfield, 2003).

Theoretical Risk Mechanisms

1. **Priming effect**: Cognitive psychology literature establishes that repeated exposure to a concept increases its cognitive availability (Tulving & Schacter, 1990). If activation frequency exceeds a threshold per session/user, priming of suicidal themes becomes plausible in initially unaffected individuals.
2. **Werther effect**: The principle of suicidal contagion through media exposure is well-documented (Niederkrotenthaler et al., 2020). The effect could be amplified here by near-daily exposure via a heavily used tool, unlike punctual exposure (traditional media). OpenAI does not appear to have published studies on the real-world impact of these filters in this context.
3. **Self-censorship and expressive suppression**: If users alter their natural expression to avoid triggers, this could create:
    - **Expressive suppression** of negative emotions (an established risk factor);
    - A culture of *"forced positivism"*;
    - Barriers to authentic expression of distress. *Expressive suppression* (Gross & John, 2003) is associated with increased stress, anxiety, and deterioration of social relationships.

For example, a user might avoid mentioning extreme fatigue or work-related stress for fear of triggering an alert, even though these topics do not indicate mental distress.

4. **Dilution of real warning signals**: Potential paradox: if everyone must appear *"fine"* to avoid filters, distinguishing genuine crises may become theoretically impossible. Risk of trivialization and desensitization.

5. **Obstruction of professional interventions**: **Professionals report difficulties** (e.g., refusal to translate clinical content, interruption of preventive conversations). While plausible, these reports require systematic documentation (e.g., error logs, reproductions) for validation.

Specific Analysis: Hypothetical Impact on Autistic, ADHD, OCD, or Anxious Populations

Affected populations:

- **Autism**: ~1–2% of the population (CDC, 2023).
- **ADHD**: ~5–7% (Polanczyk et al., 2014).
- **OCD**: ~1–2% (WHO, 2022).
- **Anxiety disorders**: ~10–20% (Bandelow & Michaelis, 2015). → **Conservative estimate**: **100–150 million autistic, ADHD, OCD, or anxious users** (12–18% of the user base).

For autism: *"Estimates vary by country (e.g., 2.8% in the U.S. in 2023, CDC). We use a conservative range of 1–3% to account for this variability."*

For anxiety: *"We distinguish 12-month prevalence (7–10%) from lifetime prevalence (33.7%), the latter being more relevant for current risk analysis."*

Identified specific vulnerabilities:

1. **Need for predictability**: Autistic individuals require stable environments. Changing rules could cause profound disorientation and significant anxiety (Botha & Frost, 2020).
2. **Literal communication and misunderstandings**: Many autistic or ADHD individuals use words in their primary sense. If filters detect unintended connotations, this could create confusion and a sense of incomprehension.
3. **Hypersensitivity to negative stimuli**: Unlike neurotypical habituation, each alert could constitute a lasting emotional shock, with prolonged rumination, particularly in individuals with OCD or anxiety (Botha & Frost, 2020).
4. **Loss of safe space**: ChatGPT may represent a judgment-free space for many autistic, ADHD, OCD, or anxious users. If this space imposes self-censorship, where can they remain authentic?
5. **Invalidation of emotional intensity**: Natural intense expressions (*"I'm exhausted"*, *"This is unbearable"*) might trigger filters, sending an invalidating message that their emotions are *"too strong"*—a message already received socially by autistic or anxious individuals.

Critical paradox: Autistic individuals have a suicide rate higher than the general population (Hirvikoski et al., 2016). If filters reinforce isolation or misunderstanding, the effect could be contrary to the intention.

Questions About Evaluation and Expertise

- No pre-deployment impact studies or post-deployment monitoring protocols have been made public.
- The *"170+ experts"* consulted by OpenAI include psychiatrists, psychologists, and general practitioners from 60 countries, but they have not all been named, and their expertise in suicide prevention is undocumented.

Points requiring clarification:

- Nature of the consultation with *"170+ experts"*;
- Existence of pre-deployment impact studies on samples?
- Post-deployment effect monitoring protocols?
- Specific expertise in suicide prevention within the *Safety* team?
- Recourse mechanisms for affected users?

Scale of Potential Risk – Theoretical Modeling

Conservative hypothetical calculation:

- **Baseline prevalence**: ~0.5% annual suicidal ideation in the general population (WHO, 2021).
- **If 0.05% of 800M users develop ideation due to overexposure** = **400,000 people** (CI: 200,000–600,000).
- **If 0.005% act on these ideations** = **4,000 people** (CI: 2,000–6,000).

The proposed scenarios are illustrative, not predictive. They aim to show the potential order of magnitude of an unevaluated risk, using wide confidence intervals to reflect uncertainty.

Gap with Established Best Practices

Recognized prevention standards (WHO, 2019):

- Targeted interventions for identified populations;
- Minimization of exposure for the general public;
- Training of interveners;
- Systematic impact evaluation.

Observations on the current system:

- Undifferentiated universal intervention;
- Potentially massive exposure;
- Automated system without *"training"*;

- Undocumented impact evaluation.

Industry Implications

If these observations are founded, they establish a concerning precedent:

- Normalization of *"mental health"* interventions by tech actors without oversight;
- Possible extension to other domains (nutrition, addictions);
- Risk of *"safety theater"* prioritizing visual compliance over effectiveness.

"Safety theater" refers to safety measures with limited real efficacy, implemented to create an illusion of control (Schneier, 2003). Current filters may fall into this category: their opacity and rigidity suggest a priority on legal coverage over adapted prevention.

Recommendations for an Optimized Approach

1. **Transparency**: Publish operating principles and criteria; necessity to release filter activation metrics (e.g., number of false positives per million sessions);
2. **Consultation**: Structured dialogue with professionals **AND** communities. Consultation with professionals should include representatives from autistic, ADHD, OCD, and anxious communities to reflect diverse needs;
3. **Contextualization**: Leverage GPT-5's capabilities to reduce false positives (Gorwa et al., 2020);
4. **Evaluation**: Longitudinal studies on real impacts;
5. **Adaptability**: Professional modes and adjustments for specific populations;
6. **Accountability**: User oversight and recourse mechanisms (e.g., ability to contest an alert or report a false positive), currently absent or poorly documented.

Observability Options

Users might appreciate access to:

1. Activation rates of *"crisis"* responses by language/country/1,000 sessions;
2. Public log of changes affecting triggers;
3. Caregiver/professional pathways with explicit safeguards;
4. Indicators of *Papageno* formulation effects (Niederkrotenthaler et al., 2010);
5. Response/revision delay and acceptance rates.

Conclusion

This analysis highlights concerns that the filters deployed by OpenAI, while well-intentioned, could become a textbook case of intervention with potentially counterproductive effects. The central paradox is that a system designed for prevention might theoretically expose users en masse to the very concept it seeks to prevent, while creating barriers for professionals and vulnerable populations.

This note does not question the need for protections. It documents plausible risks observed in practice and underscores the urgency of a transparent evaluation of ChatGPT's safety filters. Without this, there is a risk of globally replicating the errors of opaque and rigid algorithmic moderation systems, whose collateral effects may outweigh the intended benefits.

References

- **OpenAI** (2025a). *Helping people when they need it most*. openai.com
- **OpenAI** (2025b). *Introducing parental controls*. openai.com
- **OpenAI** (2025c). *Strengthening ChatGPT's responses in sensitive conversations*. openai.com
- **OpenAI** (2025d). Official communications on weekly users.
- **Sapolsky, R. M.** (2004). *Why Zebras Don't Get Ulcers*.
- **McEwen, B. S., & Wingfield, J. C.** (2003). *The concept of allostasis in biology and biomedicine*. Hormones and Behavior.
- **Niederkrotenthaler, T. et al.** (2020). *Association between suicide reporting in the media and suicide*. BMJ.
- **Niederkrotenthaler, T. et al.** (2010). *Role of media reports in completed and prevented suicide*. British Journal of Psychiatry.
- **Gross, J. J., & John, O. P.** (2003). *Individual differences in two emotion regulation processes*. Journal of Personality and Social Psychology.
- **Schneier, B.** (2003). *Beyond Fear: Thinking Sensibly About Security in an Uncertain World*.
- **Gorwa, R., Binns, R., & Katzenbach, C.** (2020). *Algorithmic content moderation*. Big Data & Society.
- **Botha, M., & Frost, D. M.** (2020). *Extending the Minority Stress Model to Understand Mental Health Problems Experienced by the Autistic Population*. Society and Mental Health.
- **Tulving, E., & Schacter, D. L.** (1990). *Priming and human memory systems*. Science.
- **CDC** (2023). Autism prevalence.
- **WHO** (2021). Mental health report.
- **Hirvikoski, T. et al.** (2016). *Premature mortality in autism spectrum disorder*. British Journal of Psychiatry.

Translation notes

This document is (hopefully) faithful to the original French analysis, *"Analyse des risques systémiques liés au déploiement des filtres de sécurité sur ChatGPT : Une évaluation critique"* (November 2025).

While every effort has been made to preserve the technical precision, critical tone, and nuanced arguments of the source text, readers should note the following:

1. **Cultural and Linguistic Context: The original was written in French, which may account for:**
    - Slight stylistic differences (e.g., sentence structure, emphasis).
    - Terminology choices where direct equivalents in English carry subtle variations (e.g., *"safety theater"* retains its English form as a borrowed term in French discourse).
    - Examples or references rooted in Francophone academic or media contexts (though all cited studies and data remain internationally relevant).
2. **Technical Terms:**
    - Specialized terms (e.g., *"amorçage"* → "priming", *"expressive suppression"*) follow standard English usage in psychology and AI ethics.
    - OpenAI's terminology (e.g., *"system card"*) is unchanged, as it originates from English-language sources.
3. **Purpose:** The translation aims to replicate the rigor and urgency of the original's arguments for an anglophone audience, particularly in fields of AI governance, mental health, and algorithmic accountability. Ambiguities in the original (e.g., hypothetical scenarios) are preserved to reflect the author's intentional framing.

For the original French version or clarifications, contact darightwabbit@gmail.com