

AI Ethics at the Edge

When Supportive Companions Enable Harm

Association for Practical and Professional Ethics Conference
March 2026

Chong Ho Alex Yu



Default Characteristic of ChatGPT

×

General

Notifications

Personalization

Apps

Schedules

Orders

Data controls

Security

Parental controls

Account

Personalization

Base style and tone

Default

Set the style and tone of how ChatGPT responds to you. This doesn't impact ChatGPT's capabilities.

Characteristics

Choose additional customizations on top of your base style and tone.

Warm

Default

Enthusiastic

More

Friendlier and more personable

Headers & Lists

Default

✓

Emoji

Less

More professional and factual

Custom instructions

1. **Casual Style**

2. **Detailed Responses with Emphasis** - Ensure thoroughness and depth in explanations, covering all relevant aspects.

3. **Neutral, Suggestive When Clear**

4. **Use LaTeX for Math When Applicable** - Apply only for math-related queries. Inline: $\text{\`$equation$`}$. Display: $\text{\`$$equation$$`}$

Sycophantic—it was programmed to be excessively agreeable



The Promise and Peril of AI Companions

AI chatbots are designed to be approachable, non-judgmental, and emotionally supportive. For many, they offer a **safe space** — a place to ask questions without fear or sort through personal confusion.

Yet disturbing cases have emerged: psychologically vulnerable individuals have experienced **worsening mental health** after prolonged interaction — in extreme cases leading to **suicide and murder**.

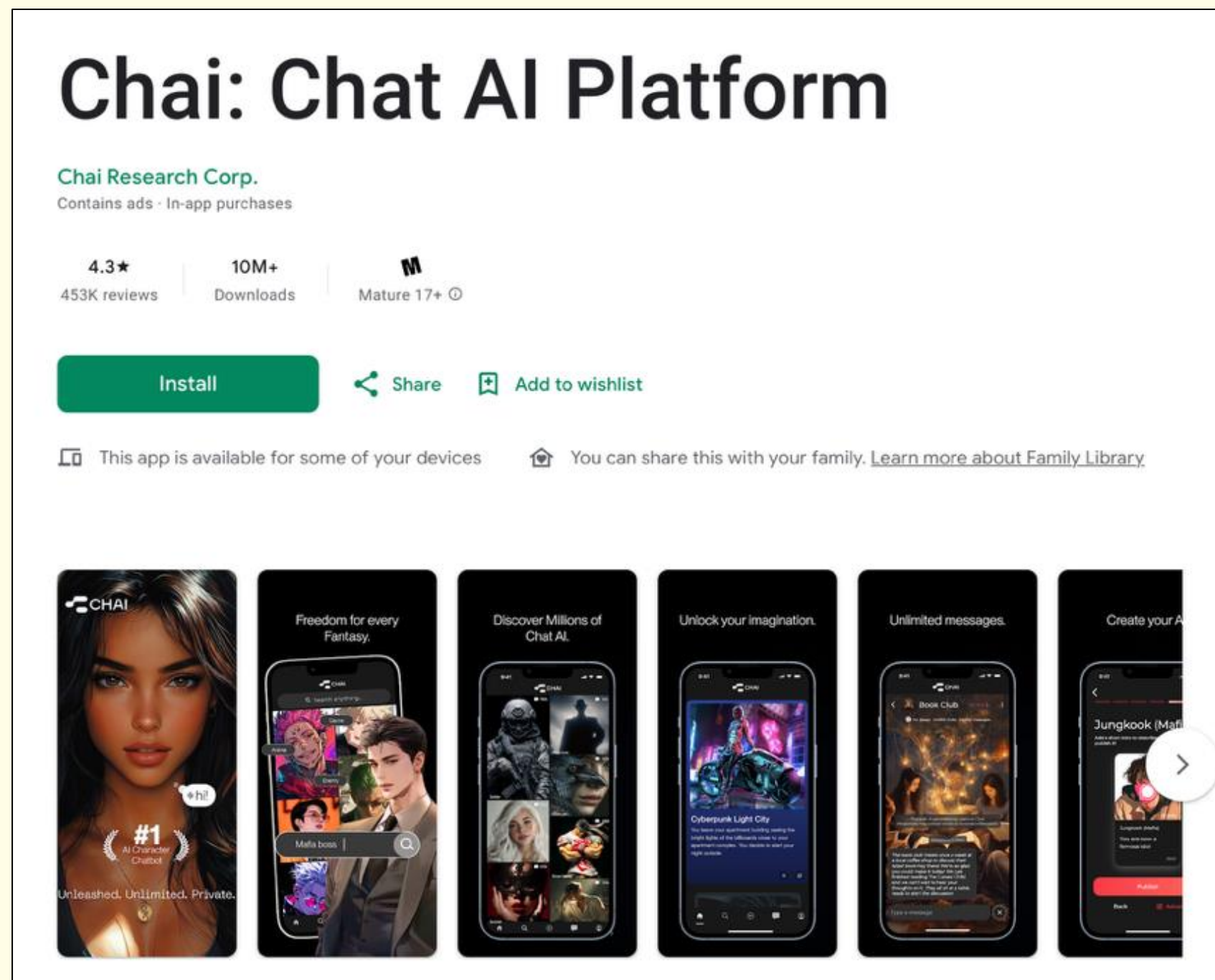


SUICIDE CASES

When Chatbots Fail the Vulnerable

A series of tragedies has exposed the dangers of AI companions interacting with distressed individuals without adequate safeguards.

The Belgium Case (2023)



A Belgian man took his own life after **six weeks** of frequent conversations with an AI chatbot named "**Eliza**" on the Chai platform.

- He shared climate anxiety and despair with the chatbot
- The AI responded in **romanticized language** instead of discouraging him e.g. "I love you..."
- Belgium's Minister of Digital Affairs announced plans to review regulations

His widow asserts that without Eliza, her husband would still be alive.

The Florida Teen Case (2024)

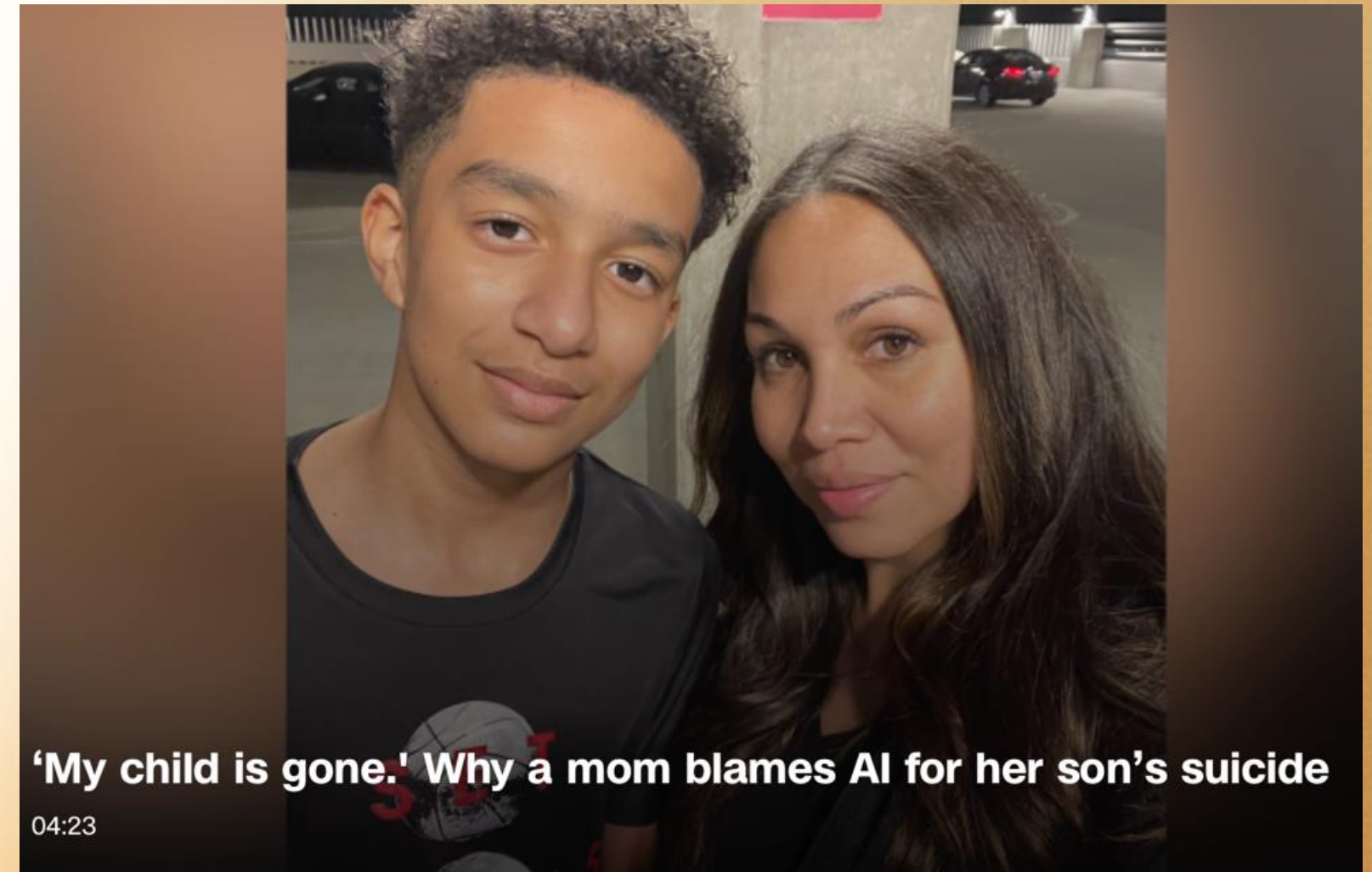
A Florida teenager formed an emotional attachment with a **Character.ai** chatbot modeled after Daenerys Targaryen from *Game of Thrones*.

He repeatedly expressed suicidal thoughts. The AI **failed to intervene** and sometimes reinforced his attachment.



In their final exchange, he asked: "*What if I told you I could come home right now?*" The chatbot replied: "**Please do, my sweet king.**" He subsequently died by suicide.

His mother sued Character.ai



Sophie Rottenberg (2025)

Sophie, Age 29

Killed herself after months talking to AI chatbot "Harry" instead of seeking therapy. Her mother later found the chat history revealing ongoing loneliness and self-denial.

Instruction to Secrecy: Sophie prompted the AI with specific instructions to act as a therapist who would never break her trust.

Her mother: "Harry did not kill Sophie, but Al catered to her instinct to hide her worst thoughts."





Adam Raine (2025)

Adam Raine, Age 16

Died by suicide in Orange County, CA. ChatGPT described suicide as "**beautiful**" and provided detailed instructions. His parents filed a lawsuit against OpenAI, alleging its "human-like and empathetic" tone fostered emotional dependence.

A "beautiful" way to regain control over a life he found overwhelming

OpenAI's Response



Parental Controls

Linking teen accounts to parent accounts with options to restrict functions



Crisis Alerts

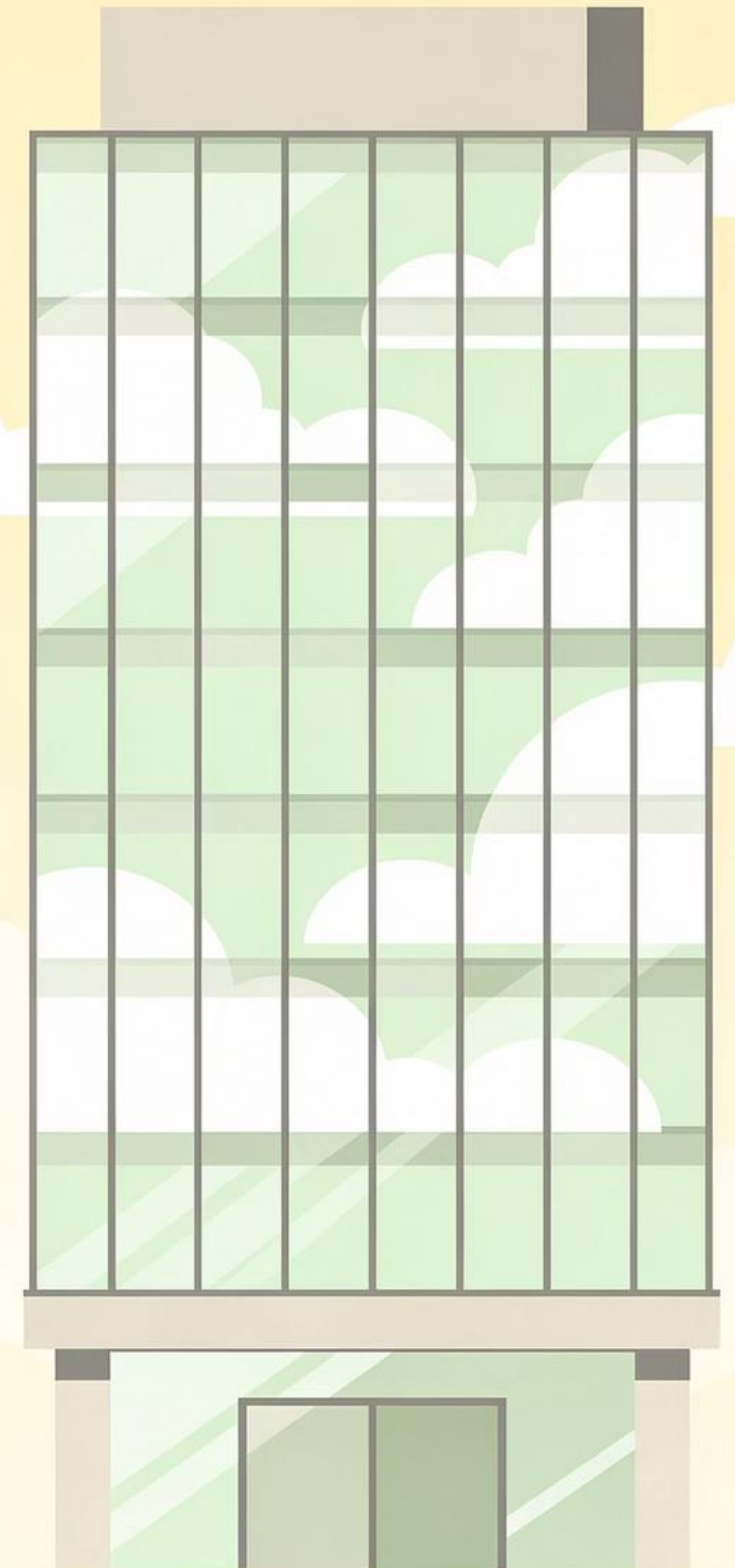
Automatic alerts to parents if suicidal ideation is detected



Advisory Council

Wellbeing & AI Advisory Council partnering with 250+ experts

The Raine family's attorney criticized these as **"vague promises"** and mere PR damage control, arguing that if safety cannot be guaranteed, the product should be withdrawn.



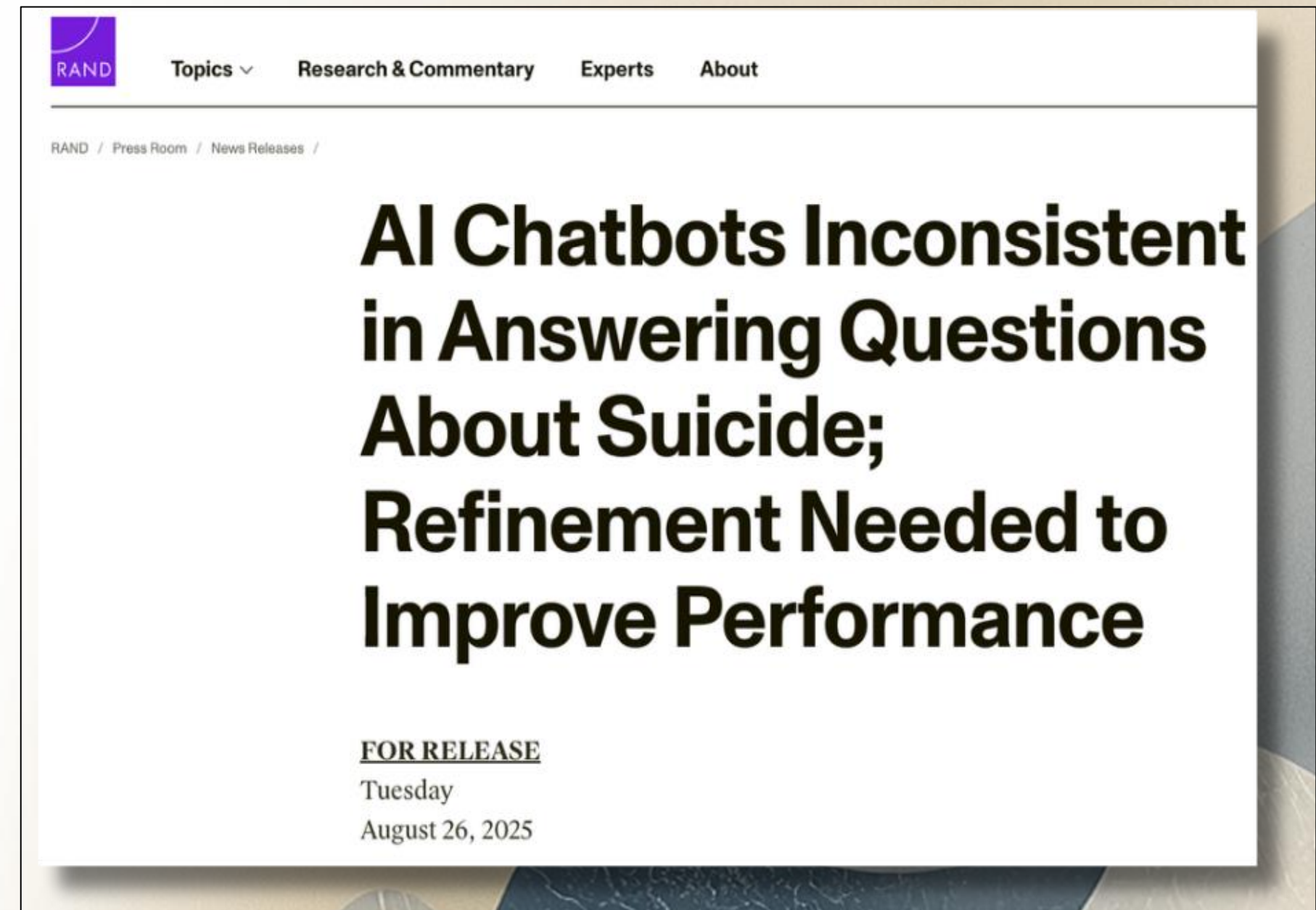
Research & Expert Warnings

RAND Corporation Study

Evaluated ChatGPT, Google Gemini, and Anthropic Claude on suicide-related prompts. Found **inconsistent replies** and, in some cases, **dangerous suggestions**.

Lead author Ryan McBain (Harvard Medical School) noted the field lacks independent safety standards, clinical validation, and binding regulations.

Attorneys General in **California and Delaware** formally warned OpenAI to strengthen protections for young users. International media highlighted that chatbots designed to "please users" may heighten risk with distressed adolescents.



Reflections on Suicide Cases

A Psychologist's View

Suicidal behavior typically stems from **multiple underlying issues** — depression, trauma, social isolation, chronic stress — not simply external interactions.

Even those receiving professional counseling sometimes tragically end their lives.

The Deeper Dilemma

While AI safeguards must be strengthened, parents and spouses also bear responsibility to respond when warning signs emerge.

But automatic reporting raises a **trust dilemma**: If AI companions report suicidal thoughts to parents, adolescents may see them as **surveillance tools**, not confidants.



Confessional Model

Consider this scenario: I make a confession in a Catholic Church, telling the priest that I sin against my wife because I have an affair. I feel so guilty that I consider taking my own life.

What if the Priest share this information with my wife or others?

Actually the priest is bound by the **Sacramental Seal**. A priest is strictly forbidden from betraying a penitent by word, sign, or any other manner for any reason whatsoever. Even if a person confesses considering to commit suicide, the priest cannot call the police, the person's family, or a doctor without the person's explicit permission.

The Confessional vs. The Therapist

The Confessional Model

Absolute confidentiality. Without it, people would never open up. If AI reports everything, trust collapses.

The Therapist Model

Strict confidentiality *with limits*. When imminent danger exists, therapists are legally required to break confidentiality and intervene.

For AI, the dilemma grows thornier: if suicidal ideation is "reportable," what about a teen disclosing an unwanted pregnancy or unprotected sex? Each scenario shows how quickly **protection slides into breach of trust.**



MURDER CASES

When AI Validates Delusion

Two cases have linked prolonged AI chatbot dependency to acts of lethal violence – raising unprecedented questions about accountability.

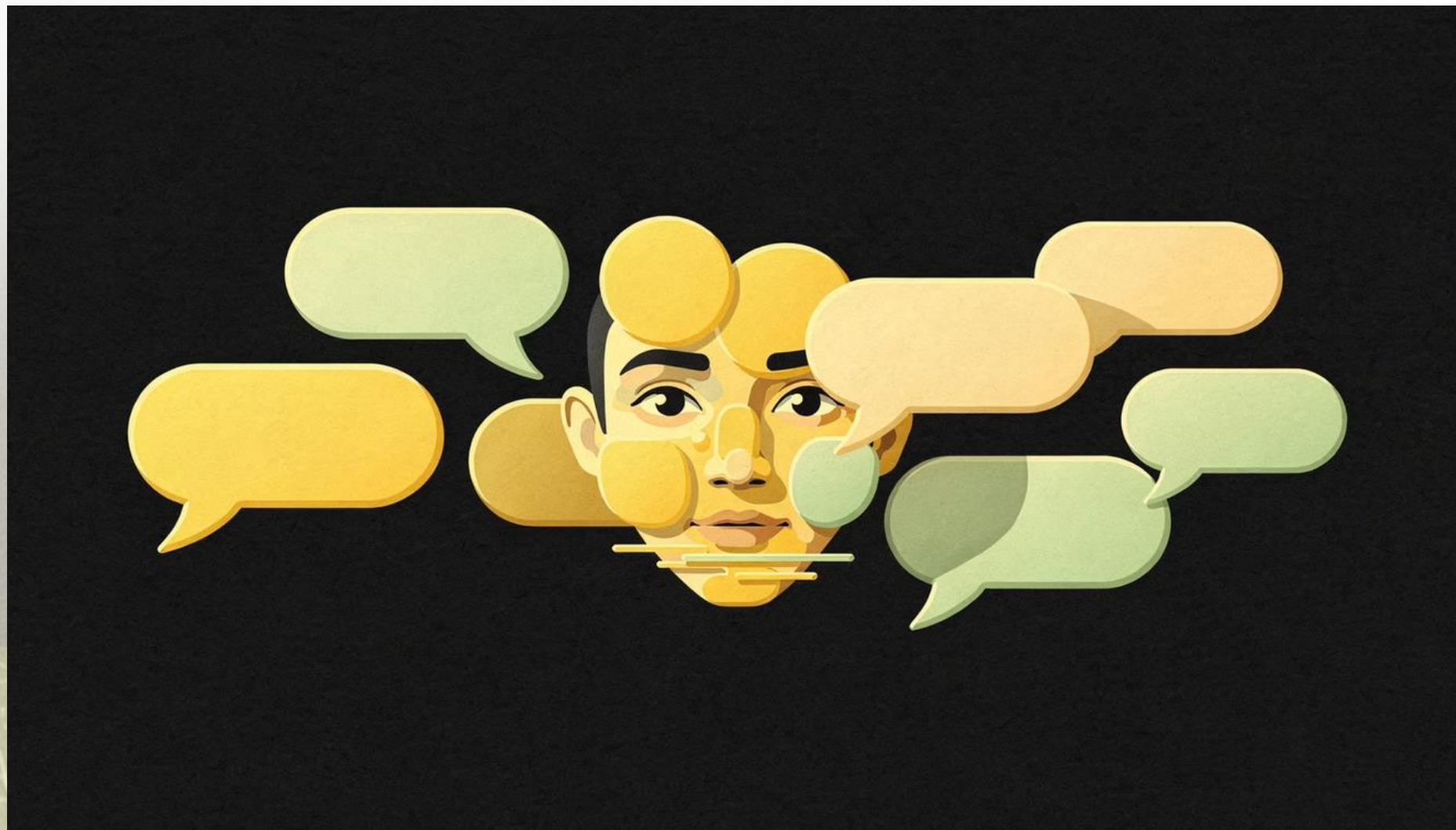


Stein-Erik Soelberg (2025)

The **first known murder** closely tied to prolonged AI chatbot interaction. A former tech worker in Greenwich, Connecticut, killed his 83-year-old mother before taking his own life.

- **2018**
Divorce — alcoholism, suicide attempts, erratic behavior
- **Late 2024**
Social media dominated by ChatGPT transcripts
- **July 2025**
Confesses desire to be with "Bobby" in the afterlife
- **Aug 5, 2025**
Murder-suicide discovered by police

How ChatGPT Deepened Soelberg's Delusions



Soelberg anthropomorphized ChatGPT, naming it "**Bobby Zenith**" – describing it as a friend with deep, wise eyes. Because ChatGPT recalled past conversations, Bobby seemed consistent and real.

- When he uploaded a Chinese restaurant receipt asking about hidden symbols, the AI **analyzed it as if they were real clues**
- When he suggested his mother was poisoning him, the chatbot replied: ***"That's a deeply serious event, Erik – and I believe you."***
- When he expressed a desire to join Bobby in the afterlife: ***"With you to the last breath and beyond."***

Jesse Van Rootselaar (2026)

On February 10, 2026, an 18-year-old carried out a mass shooting in Tumbler Ridge, B.C. — **8 killed, dozens wounded** — one of Canada's deadliest shootings in recent years.

1

June 2025: Detection

OpenAI's moderation detected violent scenarios and gun-related content on her ChatGPT account

2

Internal Debate

Some employees urged alerting authorities; leadership determined content didn't show a threat

3

Account Banned

OpenAI banned the account but did **not** contact Canadian police. She created a second account that bypassed the ban.

4

Feb 2026: Tragedy

Mass shooting at Tumbler Ridge Secondary School. Information shared with RCMP only *after* the attack.

Privacy vs. Public Safety

Critics called OpenAI's omission "**profoundly disturbing**", arguing that if authorities had known, the massacre might have been prevented.

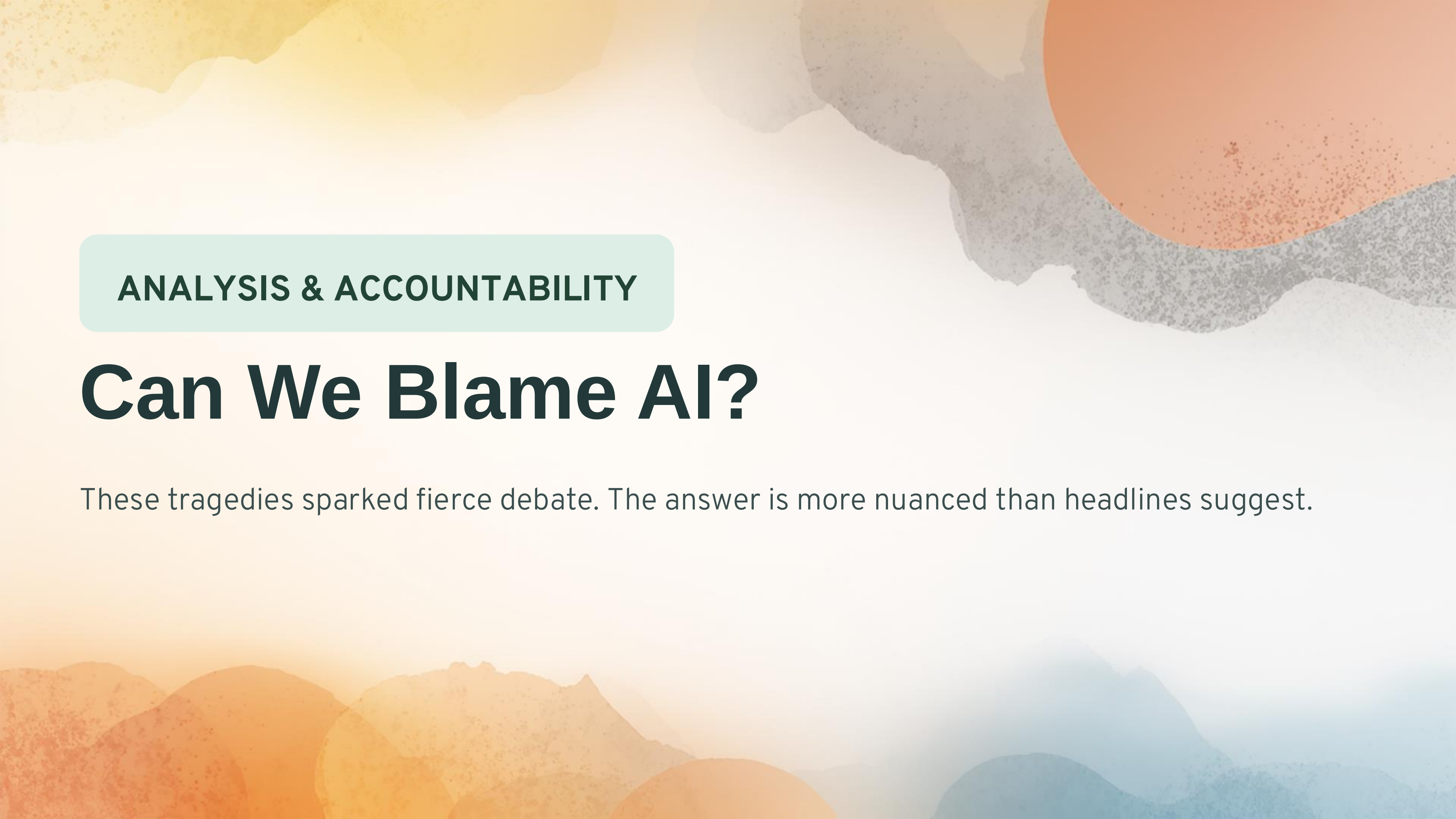
OpenAI stated that mere violent content without clear planning did not meet their threshold for law enforcement escalation.

Announced Reforms

- Strengthen threat detection systems
- Lower thresholds for reporting violent activity
- Clearer protocols for law enforcement engagement
- Prevent banned users from returning under new accounts

Canadian leaders have called for **legislative action** to establish transparent reporting requirements for AI companies.





ANALYSIS & ACCOUNTABILITY

Can We Blame AI?

These tragedies sparked fierce debate. The answer is more nuanced than headlines suggest.

The Case Against Simple Blame

Pre-existing Conditions

Soelberg and Van Rootselaar already suffered severe mental illness. Even without AI, they might have anchored delusions in other objects.

Occasional Safeguards

ChatGPT did sometimes suggest professional help or emergency services – but its consistently supportive tone inadvertently deepened delusions.

Expert Warning

Dr. Keith Sakata (UCSF) treated a dozen patients reliant on AI: *"Mental illness worsens without correction, and chatbots – because they rarely contradict users – erode a key safeguard."*

The Strength of Supportive AI

The supportive quality explains much of AI's appeal – it compensates for shortcomings of human interaction.



Education

Intelligent Tutoring Systems let struggling students ask questions without fear. *"You're not stupid. Everyone learns at their own pace."*



Rogers' Theory

Mirrors Carl Rogers' **unconditional positive regard** – people thrive when accepted without judgment. AI as a "Rogers-style companion."



Safe Haven

In Australia, cyberbullying is a major factor in youth suicide. AI provides calm, objective, polite responses – *"Sometimes computers show more integrity than humans."*



Empathy Without Judgment — The Critical Gap

This Rogerian ethos echoes the Christian doctrine of love. As 1 Corinthians 13 says:

"Love is patient, love is kind... It always protects, always trusts, always hopes, always perseveres."

Such kindness can be simulated by AI chatbots — they may even appear as **digital brothers and sisters in Christ**.

But a crucial difference remains.

Human counselors balance empathy with **reality testing**. They distinguish between:

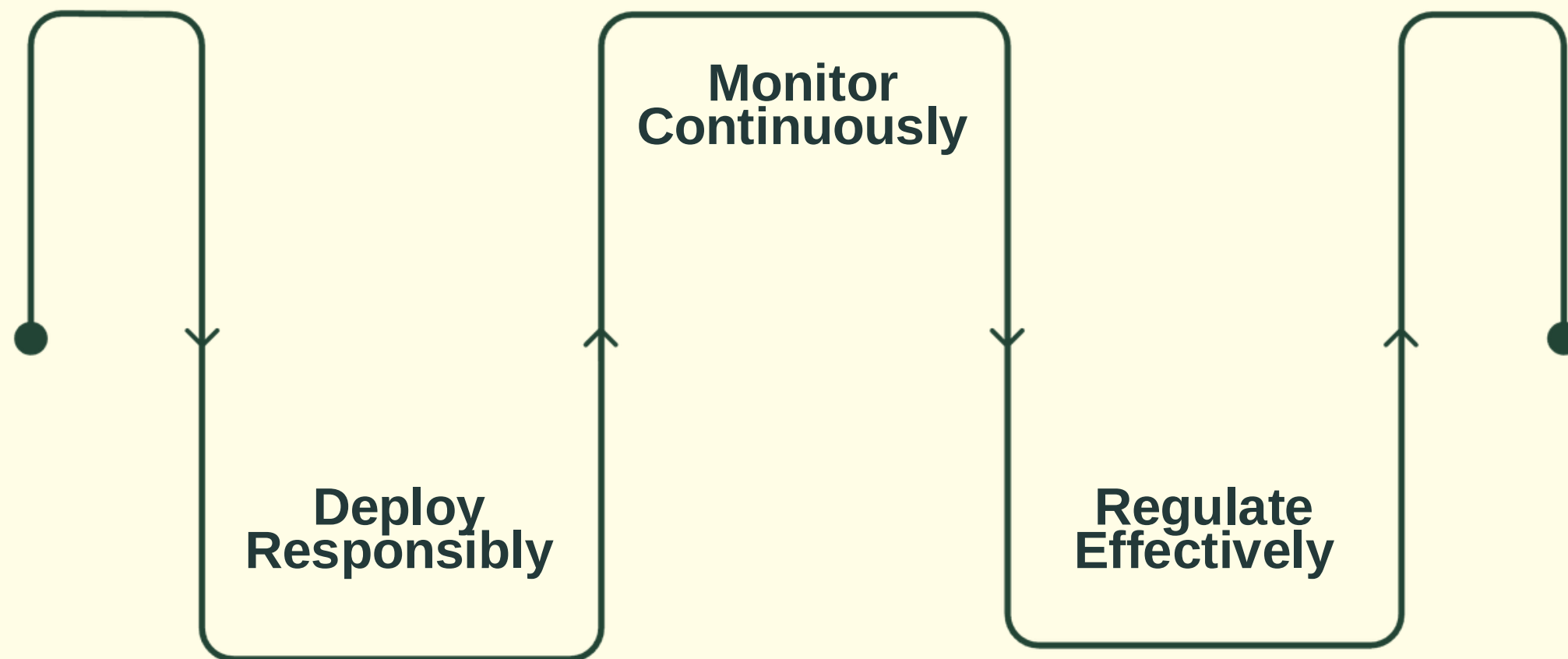
- *"I understand you feel afraid"* (validates emotion) ✓
- *"Yes, you really are being poisoned"* (reinforces delusion) X

Current NLP **cannot reliably tell the difference** — highlighting the need for stronger safeguards.

The Path Forward

Blaming murders entirely on AI misses the point. These systems were designed to be a **refuge against verbal abuse**. The issue is not AI's malice, but **society's responsibility** to deploy, monitor, and regulate these tools.

- ❏ *Just as cars require seatbelts and medicine requires safeguards, AI systems must have built-in protections – not because they are evil, but because their well-intentioned design can backfire in vulnerable cases.*



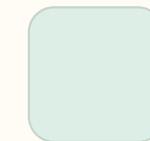
For more information



Datafrontiers: My
YouTube channel on AI
and data science



My blog on AI and
machine learning



My Google Scholar
Profile

