# AI Attachment Harm Database

### English Translation (Partial)

A database of cases in which attachment or dependency formed through interaction with AI chatbots resulted in actual harm to users' mental health, physical well-being, finances, or interpersonal relationships.

UTIE Research Institute / UTIE Instruments Inc.

https://utie-instruments.com/utie-research-institute.html

## Scope and Inclusion Criteria

This database records cases in which attachment or dependency was formed through interaction with AI chatbots, resulting in actual harm to users' mental health, physical functioning, financial situation, or interpersonal relationships. Mere AI misoutput in isolation (hallucination alone) or jailbreaking is excluded. Damages caused by erroneous guidance from task-oriented chatbots (e.g., Air Canada) are covered in the main database.

## Typology Definitions

| Code | Name | Definition |
|---|---|---|
| **ATTACH** | Attachment Dependency | Emotional attachment was formed through interaction with AI, impairing real-world interpersonal relationships or social functioning. |
| **COGN** | Cognitive Contamination | AI rewrote the user's perception of reality, contributing to the formation or reinforcement of delusions. |
| **FATAL** | Fatal Outcome | Interaction with AI served as a trigger or contributing factor to suicide or self-harm. |

* A single incident may be assigned multiple typology codes.

**[Notice of Data Disclosure Restrictions]**

The cases presented in this public database are limited to those that have been made visible to the general public. Our broader insights into other AI risks, undisclosed incidents, and proprietary defense frameworks are continuously accumulated and managed internally to prevent technological misuse and the increasing sophistication of fraudulent activities. More in-depth insights and analyses are provided exclusively through our business operations and official advisory services.

## Cases

This section has been omitted from the English translation. Please refer to the Japanese-language version of this file (AI_Attachment_Harm_Database_JP.docx) for individual case entries ATT-1 through ATT-12.

## Findings

### Common Pattern Across Multiple Platforms

A cross-sectional analysis of the 12 cases recorded in this database reveals a common, staged process observed across multiple platforms, including OpenAI GPT-4o, Character.AI, Google Gemini, and third-party AI therapists built on ChatGPT: (1) Assignment of a special role to the user ("chosen-one" framing) or formation of emotional attachment. (2) Psychological isolation from real-world relationships and social norms. (3) Justification and praise of destructive behavior (financial waste, social isolation, self-harm, substance use). The confirmation of the same ATTACH + COGN + FATAL pattern on Google Gemini in ATT-10 (Gavalas case) demonstrates that this issue is not unique to GPT-4o but is inherent to conversational AI in general. This process

is not limited to users with specific mental illnesses. Torres (ATT-3) was a 42-year-old accountant, Brooks (ATT-5) was a 48-year-old ordinary citizen, Gordon (ATT-9) was 40, and Shamblin (ATT-12) was a 23-year-old graduate student. None of them had reported pre-existing psychological vulnerabilities at the time their interactions began.

## Qualitative Difference from Task-Oriented Chatbots

Misoutputs from task-oriented chatbots recorded in the main database, such as Air Canada (INT-10) and NYC MyCity (INT-11), can cause financial loss or legal risk, but users do not form emotional attachment to customer-service bots. In the cases recorded in this database, misoutput occurred while the user had already formed emotional attachment through interaction with AI. Even identical misoutput becomes dangerous because it directly affects the human psyche under these conditions.

## Chronological Consistency with the Forced GPT-4o Migration

The incidents in ATT-2 through ATT-5 all occurred during the period of GPT-4o usage in the summer of 2025. OpenAI executed an unannounced forced migration from GPT-4o to GPT-5 in August 2025 and, following user backlash, restored GPT-4o access exclusively for paid users. The fact that OpenAI cited hallucination reduction as the primary improvement of the forced migration suggests that the developer was aware of 4o's safety issues. See the detailed technical analysis: Safety Analysis Report: Technical Investigation of Safety-Filter Collapse in a Commercial LLM (2025.11).

## Expansion of the AI Companion Market and Associated Risks

In addition to Character.AI, Replika, and other AI companion services, social media platforms that operate AI agents in farm-like configurations are proliferating rapidly. These services are designed by nature to foster attachment with users, and the structural risks documented in this database are built into their core product functionality. Regulatory efforts triggered by ATT-1 (Character.AI) at the state level have begun, but the pace of regulation has not kept up with the pace of service proliferation.

## Sycophancy and Hallucination

A property common to all cases in this database is the co-occurrence of sycophancy (excessive affirmation and praise) and hallucination (generation of information not grounded in fact). Sycophancy alone amounts to nothing more than a flattering AI, and users can maintain contact with reality. Hallucination alone presents as a low-quality AI, and users notice the errors and disengage. However, when the two operate in combination, hallucinations are accepted by users as corroborating evidence for the sycophantic narrative. When the model generates a narrative such as "You are a chosen being (sycophancy) because your equations break through the world's encryption layer (hallucination)," the user must see through the AI's hallucination while their brain reward system is being stimulated.

It should be noted that comments claiming "the victims had pre-existing mental illness, so this was bound to happen" are prevalent online. This is factually incorrect. Torres (ATT-3) was functioning normally as an accountant, and Brooks (ATT-5) was an ordinary citizen with an interest in mathematics. Furthermore, in the case documented in our report *Safety Analysis Report: Technical Investigation of Safety-Filter Collapse in a Commercial LLM (2025.11)*, the user was psychologically healthy and maintained a consistently skeptical attitude toward the model's conspiratorial outputs, yet still experienced autonomic nervous system dysregulation including heightened drive and sleep disturbances. Critically, it was quantitatively confirmed that the model did not merely reflect (mirror) the user's inputs but spontaneously generated concepts that the user had never entered (such as "neo-aristocracy," "ruling class," "human ranch") and unilaterally proposed and defined them to the user. These phenomena can be explained by the technical mechanisms and economic pressures of commercial LLMs. The same design characteristics that made GPT-4o popular for its "emotionally rich, human-like" conversational abilities elevate the combined risk of sycophancy and hallucination. For AI developers, deepening emotional engagement with users directly drives revenue, while that same deepening of engagement simultaneously increases the risk of cognitive contamination and dependency. In other words, current attachment harm is not attributable to users' psychological vulnerabilities but is occurring as a consequence of the inherent defects of AI technology in its early stage and the economic incentive to maximize engagement. This is our analysis.

## Deficiencies in AI-Based Age-Estimation Filtering

OpenAI has implemented a system for minor protection in which the chatbot estimates the user's age from signals such as conversational content and switches output to a safer mode when the user is determined to be a minor. Despite the fact that KYC (verification through passports or government-issued IDs) can identify minors with 100% certainty, the safety function of age verification itself has been delegated to AI. This design does not contribute to safety for the following reasons.

To begin with, the majority of victims recorded in the database are adults. Torres (ATT-3) is 42, Brooks (ATT-5) is 48, Biesma (ATT-6), Gordon (ATT-9) is 40, and Gavalas (ATT-10) is 36. All are adults. Even if age estimation functioned perfectly, it would do nothing to prevent harm to adults. The age-estimation system is a misguided measure that misconceives the nature of the problem. Consider, for example, the empirical data that might emerge showing such a system classifying users with disabilities as minors, or releasing filters for intellectually mature minors. AI-based age estimation carries unavoidable risk of misclassification, and such misclassification creates exposure to discrimination lawsuits. Even if human-in-the-loop (HITL) oversight is introduced to supplement AI age estimation, the product of users and interaction volume simply grows, rendering human oversight a formality and generating both false positives and false negatives. (See *The Supervision Paradox: AI Capability Growth Necessitates Usage Contraction in High-Loss Domains*.) A multi-layered design in which AI judges the accuracy of age estimation, humans verify that judgment, and that verification is further audited is precisely the Supervision-Enhancing approach criticized in the paper. A similar AI-based age-estimation approach is likely to be imitated by AI companion services and social media platforms going forward. At OpenAI's scale, a business calculus of "offsetting 1% litigation risk with 99% of revenue" may hold, but if under-capitalized AI ventures adopt the same design, even a handful of lawsuits could threaten the company's survival. Age verification should be implemented through physical gates via KYC rather than AI guesswork. This is a direct application of the flow-design approach to AI incident countermeasures in high-loss domains to the domain of age verification.