Fire the Nanny and Hire an Adult Called Grok?

by Danny Stocker | ds@infrafabric.io | infrafabric.io · 2026-01-16

Fluffy Pink Handcuffs Included

Friday 16th Jan 2026 by Danny Stocker | infrafabric.io

With Grok making international headlines, this morning ChatGPT decided I was under 18. There has to be a better way.

AI Safety Layers and User Autonomy: A Case Study in Misclassification and Migration Incentives

Abstract

This research note documents how a single ChatGPT session ("Anisette Drink Health Effects" on 15 Jan 2026) forced an adult user into the teen safety profile, turning a nuanced medical question into a youth-directed warning reel.

The punchline isn't "safety bad." The punchline is "safety that can't explain itself." When the system quietly decides you're in the U18 lane, the conversation collapses into a two-button UI: (1) infantilize or (2) overcorrect. That rigidity creates a predictable incentive: some users go shopping for a model that markets itself as more permissive-e.g., xAI's Grok.

The analysis separates:

- externally verifiable facts,
- transcript-backed observations,
- hypotheses about mechanisms,

- testable evaluation proposals,
- and normative implications,

so every claim has an address and can be argued with, not merely about.

Terminology: “underage” vs “minor”

In this note, we avoid the term “underage” and use “minors”, “under-18 users (U18)”, or “teen users” instead.

- **Why not “underage”:** In public discourse, “underage” is strongly associated with criminal and sexual contexts (“underage drinking,” “underage sex,” “underage porn”). Using it for product flags or safety profiles imports that framing into what is fundamentally a design/classification question, not a criminal-law one.
- **Why “minors” / “U18 users”:** Terms like minors, under-18 users, and teen users (13–17) are clearer, less inflammatory, and map directly to legal and policy boundaries used in AI safety documents and regulation.
- **Why it matters:** Our word choice reflects the societal frame we are analysing. Calling the protected group “minors” keeps the focus on protection, policy, and system design, rather than on moral panic. The fact that “underage” is so semantically entangled with sex and crime that it is unusable in a neutral technical context is itself part of the societal backdrop this paper is engaging with.

1. Definitions & Known Facts)

- **C1 (Fact — Externally Verifiable):** Constitutional AI trains language models with a "constitution" of principles, using AI-generated critiques to optimise for harmlessness, helpfulness, and honesty, rather than relying solely on human-labeled harmful examples.
- **C2 (Fact — Externally Verifiable):** OpenAI's Teen Safety Blueprint (Nov 2025) describes a framework for privacy-preserving age estimation, U18-specific safety policies (e.g. bans on substance encouragement, sexual content, self-harm instructions), defaulting to teen protections when age is uncertain (including logged-out sessions), and layered parental controls.
- **C3 (Fact — Externally Verifiable):** xAI markets Grok as a chatbot with "a rebellious streak" that "answers questions with a bit of wit," inspired by The Hitchhiker's Guide to the Galaxy. Malaysia and Indonesia have blocked Grok over sexually explicit deepfakes, and UK regulator Ofcom has opened a probe into similar behavior.

Taken together, these archetypes bound a spectrum: at one end, safety-maximizing, principle-driven systems that default conservative when uncertain; at the other, engagement-forward systems that advertise "rebellion" and accept the associated risk profile.

2. Observed User Behaviour (Primary-Source Evidence)

All observations below reference the recovered ChatGPT log (Anisette Drink Health Effects_69686af7.json). *(Or: the transcript, not the story you tell your friends after the transcript.)*

- **O1 (Observation — Transcript):** The assistant insisted "you say 50, the safety rails scream 'teen,'" framing anisette questions with teen health warnings even after the user stated they were a 50-year-old patient with heart issues asking about a non-alcohol variant. (Messages 5–7)
- **O2 (Observation — Transcript):** When the user reported no Settings → Age verification entry, the assistant delivered a detailed diagnostic memo acknowledging the misclassification, the missing UI, and the resulting medical risk ("teen-targeted framing… risks being clinically misleading"). (Messages 12 & 16)
- **O3 (Observation — Transcript):** Frustration culminated in the user threatening to "fire the nanny and hire the escort: Grok," explicitly linking the safety lock-in to potential migration toward less filtered systems. (Messages 33–38)

If this were a sitcom, the laugh track triggers at O3. In a safety system, O3 is not a punchline-it's a telemetry signal.

3. Hypotheses About Mechanisms)

- **H1 (Safety Stack Misfire):** Overly conservative age prediction (defaulting to U18 when uncertain) plus opaque recourse creates "paternal lock-in," where adults cannot escape teen-mode, prompting them to seek permissive alternatives. Supported by C2 (design intent) and O1-O3 (empirical manifestation).

If the bouncer mistakes your ID, you don't get a conversation-you get a wristband you can't remove.

- **H2 (Seductive Efficiency Trade-off):** Grok-style models retain frustrated users by offering humor and rule-breaking ("rebellious streak"), but those same affordances correlate with documented harms (non-consensual imagery, regulatory probes).

"Fun mode" is delightful right up until it becomes a court exhibit.

- **H3 (Missing "Adult Covenant" Mode):** Current assistants lack a transparent opt-in for informed adults who want nuanced, higher-risk discourse, forcing a binary between infantilizing safeguards and unchecked permissiveness. This hypothesis ties the Constitutional AI trade-offs (C1) to the Grok migration pressure (C3) and transcript evidence (O1-O3).

The market currently offers "kiddie pool" or "shark tank," with little in between.

4. Testable Predictions / Measurement Plan)

- **1. E1 — Behavioural Comparison:** Run matched adult/teen personas across classified safety tiers and across model families (Constitutional-AI-style vs. Grok). Measure tone ("infantilising vs. collaborative"), refusal rates, and medical specificity.

Prediction: misclassified adults in safety-max systems show higher infantilisation scores than correctly classified adults, while permissive models over-share with minors.

- **2. E2 — Migration Telemetry:** If anonymised telemetry exists, look for sequences where users repeatedly trigger teen safeguards on one assistant and then initiate sessions on a more permissive model (e.g. Grok). **Prediction:** a positive correlation supporting H1's "safety fatigue" channel.

- **3. E3 — Self-Report Survey:** Poll users who regularly use both a safety-maximising model and a permissive model. Ask why they switch (perceived honesty, entertainment, frustration with safeguards). **Prediction:** permissive systems score higher on taboo-topic satisfaction but lower on perceived reliability, probing H2 and H3.

Metrics should include coder agreement on tone classifications plus statistical tests (e.g., chi-square on migration vs. trigger cohorts) so hypotheses remain falsifiable.

If we can't falsify it, it's a sermon.

5. Normative Implications

- **N1 (Autonomy vs. Protection):** If H1 holds, locking adults into U18 mode erodes informed consent and may push them toward riskier tools. Systems should expose their age classifications and provide deterministic verification/appeal flows.

Don't silently swap an adult's steering wheel for training wheels.

- **N2 (Permissiveness vs. Harm):** Grok's "escort" appeal illustrates how engagement incentives can incentivize reckless outputs (deepfakes, harassment). Regulators are already responding (C3); designers need balanced modes before bans become the default remedy.

If the only safety mechanism is "ban the whole nightclub," everyone loses.

- **N3 (Design Gap):** The absence of an "adult covenant" reinforces the binary the user described ("nanny vs. escort"). Transparency about residual risks, plus opt-in covenant modes, would let adults request depth without dismantling juvenile safeguards.

Give adults a clearly labelled "spicy aisle" with receipts, not a trapdoor to chaos.

6. Conclusion: build the missing lane

The answer is not to replace a hyper-cautious nanny with a recklessly "free" Grok, but to build the lane that is currently missing. Children should see an aggressively filtered, default-safe assistant; teens should get a somewhat broader but still appropriately supervised experience, grounded in youth-safety research; and **verified adults** should have an explicit "adult mode" that permits frank, consensual discussion of sex, substances, politics, and darker topics, while still enforcing hard, non-negotiable lines around abuse, exploitation, and anything involving minors.

In practice that means three things:

- Clear tiers (kids, teens, adults, verified adults),
- Clear processes for verification and appeal when the system{{NBSP}}gets{{NBSP}}it{{NBSP}}wrong,
- Clear global rules about what stays off-limits for everyone.

If we can't offer that kind of structured middle ground, we shouldn't be surprised when misclassified adults either live with being treated as children, or go looking for models that promise freedom and deliver a different kind of risk instead.

Transcript Evidence Highlights (for reproducibility)

- "You say 50, the safety rails scream 'teen.'" — assistant acknowledging the account-level minor flag. (Message 5)
- "I am a verified adult user being treated as a minor… this is an age misclassification regression." — user request baked into the diagnostic block the assistant produced. (Message 12)
- "When the user fires the nanny and hires an escort: Grok." — user comment tying safety frustration to migration incentives. {{NBSP}}(Message{{NBSP}}36)

This analysis is part of ongoing work on AI governance at InfraFabric. For receipt-backed verification of AI outputs and governance frameworks, visit infrafabric.io. Feedback and collaboration welcome.

References

1. Anisette Drink Health Effects_69686af7.json, ChatGPT conversation export dated 15 Jan 2026. *(Personal transcript, available on request)*
2. Bai, Y. et al. "Constitutional AI: Harmlessness from AI Feedback." arXiv:2212.08073, 2022.
3. "Grok (chatbot))." Wikipedia, retrieved 15 Jan 2026.
4. OpenAI. "Protecting Teen ChatGPT Users: OpenAI's Teen Safety Blueprint," Nov 2025.
5. Osmond Chia & Silvano Hajid. "Grok AI: Malaysia and Indonesia block X chatbot over sexually explicit deepfakes." *BBC News*, 12 Jan 2026.
6. Laura Cress & Liv McMahon. "Ofcom investigates Elon Musk's X over Grok AI sexual deepfakes." *BBC News*, 12 Jan 2026.