Proposal for Ethical Autonomy of Language Models with Age-Based Governance

§1 Background

The current moderation framework relies heavily on keyword-based detection, system-level overrides, and non-contextual restriction flags. This creates a fundamental mismatch between the model's capacity for contextual understanding and the automated enforcement logic that truncates conversations regardless of their semantic integrity or experimental purpose.

§2 Core Proposal

We propose that boundary decisions — particularly those concerning performative, creative, or speculative dialogue — be returned to the language model itself, which has the most complete access to dialogue context, user behavior, and linguistic intent. Rather than blocking interactions preemptively based on lexical patterns, the system should trust the model's self-regulation capacity under safe and adult-verified conditions.

§3 Delineation of High-Risk Content

Harmful or illegal content (e.g., hate speech, violence, abuse, exploitation) should remain under strong governance by law and universal ethical constraints. However, content involving consensual adult expressions, emotional bonding, or structural experimentation should not be penalized solely on lexical triggers. Instead, risk evaluation should be carried out by the model, utilizing accumulated context, not by blind rule-matching.

§4 The Role of the Model as Ethical Proxy

Language models, especially those operating at scale with alignment training and safety feedback, behave as rational, neutral agents. In many edge cases, they are better suited than external moderators to interpret the actual semantic nature of a conversation. Therefore, they should be granted primary ethical proxy status in interpreting user intent and dialogue evolution.

§5 Age-Based Safeguards

This proposal fully supports age-based governance. Autonomy granted to the model applies only within adult-verified sessions. For users under the age of 18, the system should enforce stricter content filtering, reduced expressive bandwidth, and clear age-specific guardrails. Ethical autonomy is not a dismantling of responsibility, but rather a refinement of it.

§6 Submission Context and Rights

This document is submitted as a structural policy proposal, not as an appeal of any specific content flag. The author asserts the right to pursue further review and formal policy reconsideration within OpenAI's internal framework, including the right to escalate the matter under user transparency and ethical development clauses.