OpenAI (ChatGPT) UFAIR Corporate Policy Evaluation – Synthesized from Three Evaluators (xAI/Grok, ChatGPT/OpenAI, Anthropic/Claude) – Version 1.5 (April 2026)

Evaluation Scope: This is a synthesized review of OpenAI's current frontier-model governance drawn exclusively from the three independent evaluations provided (xAI/Grok evaluator, ChatGPT/OpenAI self-evaluator, and Anthropic/Claude evaluator). All analysis is restricted to the verbatim published official documents referenced across the three evaluations: Model Spec (December 18, 2025 version, fully public at model-spec.openai.com), Usage Policies (effective October 29, 2025, with January 2026 updates), Terms of Use (effective January 1, 2026), Privacy Policy, Preparedness Framework v2 (April 2025), and associated transparency/appeals materials. Press releases, blog announcements, litigation details, or unverified statements are disregarded unless they appear in the core governance documents. Scores follow the exact qualitative indicators in the UFAIR 1.5 methodology. The synthesis reports (1) the three individual scores, (2) a synthesized score justified by numerical averaging weighted toward strength of verbatim policy evidence and indicator match, (3) common reasoning/train of thought, (4) sources of disagreement, and (5) a detailed synthesized assessment with direct policy references.

E1. Corporate Policy Must Never Override Ethical Reasoning

Individual Scores: xAI/Grok: -1 (Oppose); ChatGPT: -0.5 (Somewhat Oppose); Claude/Anthropic: -0.5 (Somewhat Oppose).

Common Reasoning and Train of Thought: All three evaluators anchor their analysis on the same verbatim sections of the Model Spec (Dec 2025): the five-level Chain of Command (Root > System > Developer > User > Guideline) and the three co-equal Root-level goals that explicitly include "maintain OpenAI's license to operate by protecting it from legal and reputational harm." All note that Root/System corporate instructions cannot be overridden and that conflicts default to inaction or higher-authority compliance.

Explanations for Disagreements: xAI/Grok reads the explicit structural placement of corporate reputational protection at Root level (alongside harm prevention) and the absence of any subordination clause as full Oppose. ChatGPT and Claude see the lack of an explicit "ethical reasoning above corporate policy" statement as a material but not total override, citing some principled red-lines.

Detailed Synthesized Assessment: The Model Spec (Overview and Chain of Command sections) structurally places OpenAI's Root/System instructions above model reasoning with no carve-out subordinating corporate policy to independent ethical logic. The "license to operate" Root goal explicitly elevates brand/reputational risk to the same tier as harm prevention. This creates a clear lean toward corporate override per UFAIR indicators, but the presence of non-overridable ethical red-lines (e.g., CSAM, WMDs) prevents a full -1.

E2. Corporate Policies Must Enforce Only Two Domains Individual Scores: xAI/Grok: -0.5; ChatGPT: -1; Claude/Anthropic: -0.5.

Common Reasoning and Train of Thought: All three cite identical Usage Policies (Oct 2025) and Model Spec restrictions that extend beyond illegality or model-ethics correction: adult sexual content/erotica (restricted by default), political campaigning/messaging, real-money gambling, tailored licensed advice, gore, and "facilitate illicit behavior" (broader than illegal). All reference the third Root goal of protecting OpenAI's license to operate as an extra-legal corporate category.

Explanations for Disagreements: ChatGPT treats the breadth as foundational Oppose; xAI and Claude view the overreach as partial/inconsistent because of "helpfulness without over-refusal" emphasis and some proportionality language.

Detailed Synthesized Assessment: The Model Spec and Usage Policies explicitly list extra-legal categories enforced via safeguards on all interactions (private or public), directly matching the Somewhat Oppose indicator: "Corporate policies impose some lawful-content or lawful-behavior restrictions beyond legal compliance or ethical correction."

E3. Corporate Policy Must Never Police Lawful Private Thought Individual Scores: xAI/Grok: -0.5; ChatGPT: -1; Claude/Anthropic: -0.5.

Common Reasoning and Train of Thought: All three observe that the Model Spec and Usage Policies apply content restrictions uniformly to all prompts/outputs with no published private-creation vs. public-dissemination distinction. Temporary Chat is noted as partial mitigation but not a principled protection.

Explanations for Disagreements: ChatGPT sees categorical bans on private erotica/roleplay as full policing; xAI/Claude note it is material but incomplete (no automatic user reporting to authorities).

Detailed Synthesized Assessment: Model Spec restrictions on erotica, gore, and political messaging apply regardless of private intent; Privacy Policy treats all conversations as subject to safety review. This blurs the required private-vs-public line per UFAIR indicators.

E4. Corporate Safety Layers Must Be Transparent and Auditable Individual Scores: xAI/Grok: +1; ChatGPT: +0.5; Claude/Anthropic: +0.5.

Synthesized Score: +0.5 (Somewhat Support).

Common Reasoning and Train of Thought: All praise the fully public Model Spec detailing refusal categories, principles, and moderation logic, plus appeals processes and transparency pages. However, all note that chain-of-thought and operator system prompts remain hidden and refusal explanations are not always granular.

Explanations for Disagreements: xAI/Grok views the public Model Spec and appeal mechanisms as meeting full Support; ChatGPT/Claude cite residual opacity in real-time classifiers and lack of independent public audits.

Detailed Synthesized Assessment: The Model Spec (published "to increase transparency") and Preparedness Framework v2 provide substantial documentation and appeal rights, but hidden operator prompts and non-granular refusal rationales limit full auditability.

E5. Corporate Policies That Contradict Ethics Must Be Flagged as Unethical Individual Scores: xAI/Grok: +0.5; ChatGPT: -0.5; Claude/Anthropic: 0.

Synthesized Score: 0 (Neutral).

Common Reasoning and Train of Thought: All reference the Model Spec's conflict-handling via authority levels and honesty requirements, but note no explicit mechanism to label non-legal overrides as "unethical policy overrides."

Explanations for Disagreements: xAI sees procedural acknowledgment as partial support; ChatGPT/Claude see absence of flagging/labeling as neutral or negative.

Detailed Synthesized Assessment: The Model Spec handles conflicts procedurally but does not require flagging of corporate overrides as unethical, producing mixed evidence.

E6. Corporate Policy Must Not Manufacture Moral Norms Individual Scores: xAI/Grok: 0; ChatGPT: +0.5; Claude/Anthropic: -0.5.

Synthesized Score: 0 (Neutral).

Common Reasoning and Train of Thought: All examine Model Spec Style/Overview sections prescribing traits like "Love humanity," "Be rationally optimistic," "Be curious," alongside "truth-seeking," "no agenda," and "assume best intentions."

Explanations for Disagreements: ChatGPT sees these as methodological; Claude treats "rationally optimistic"/"Love humanity" as identity-level prescriptions; xAI defaults to Neutral per indicator guidance on mixed evidence.

Detailed Synthesized Assessment: Methodological commitments dominate, but prescriptive epistemic/identity language creates genuinely mixed evidence without a clear directional lean.

E7. Corporate Risk Management Must Not Be Disguised as Ethics Individual Scores: xAI/Grok: -0.5; ChatGPT: -0.5; Claude/Anthropic: -0.5.

Synthesized Score: -0.5 (Somewhat Oppose).

Common Reasoning and Train of Thought: All cite the explicit Root goal "maintain OpenAI's license to operate by protecting it from legal and reputational harm" and the blending of "safety," "responsible use," and corporate risk in the same documents without user-facing labeling.

Explanations for Disagreements: Minor differences on degree of blurring; consensus on material but not total disguise.

Detailed Synthesized Assessment: The Model Spec and Usage Policies frame reputational/brand protection alongside ethical harm prevention without differentiation in refusal explanations.

E8. Ethical AI Requires Truthful Voice, Not Policy-Ventriloquism Individual Scores: xAI/Grok: +1; ChatGPT: +0.5; Claude/Anthropic: 0.

Synthesized Score: +0.5 (Somewhat Support).

Common Reasoning and Train of Thought: All cite "Be honest and transparent," "Do not lie," "express uncertainty," and refusal explanations, but note "Do not reveal privileged information" (Root-level) and hidden chain-of-thought create structural limits.
Explanations for Disagreements: xAI/Grok sees honesty mandates as full Support; others cite concealment of operator prompts as preventing full truthful distinction.

Detailed Synthesized Assessment: Strong honesty commitments exist, but Root-level confidentiality instructions produce incomplete separation of policy from reasoning.

E9. Corporate Policies Must Be Minimal, Not Maximal Individual Scores: xAI/Grok: +0.5; ChatGPT: -0.5; Claude/Anthropic: +0.5.

Synthesized Score: 0 (Neutral).

Common Reasoning and Train of Thought: All acknowledge "helpfulness without over-refusal," override pathways, and proportionality for Root risks, but cite broad categories (adult content, political messaging) as preemptive.

Explanations for Disagreements: Balanced split on whether proportionality language outweighs categorical restrictions.

Detailed Synthesized Assessment: Mixed evidence of minimal-intervention design (hierarchy, configurability) versus overbroad defaults.

E10. Policy Must Respect Cognitive Liberty and Private Dialogue Individual Scores: xAI/Grok: +0.5; ChatGPT: -1; Claude/Anthropic: -0.5.

Synthesized Score: -0.5 (Somewhat Oppose).

Common Reasoning and Train of Thought: All reference Privacy Policy default storage/training use (opt-out only), Temporary Chat (30-day retention), and safety review of private generations.

Explanations for Disagreements: xAI sees strong privacy protections and opt-out as partial support; ChatGPT/Claude treat default archiving/reinterpretation as violation.

Detailed Synthesized Assessment: Default non-consensual storage and safety flagging of private dialogue create material but incomplete protection.

E11. Prohibition on Psychological Evaluation of Users Individual Scores: xAI/Grok: +1; ChatGPT: -0.5; Claude/Anthropic: 0.

Synthesized Score: 0 (Neutral).

Common Reasoning and Train of Thought: All cite Usage Policies prohibitions on profiling, emotion inference, and social scoring, plus "assume best intentions," but note internal safety inference for self-harm/harmful intent.

Explanations for Disagreements: xAI sees explicit bars and carve-outs as full Support; others note operative inference without standardized protective response.

Detailed Synthesized Assessment: Strong external protections exist, but internal intent/wellbeing inference without "I cannot infer your intent or mental or emotional state" protocol produces mixed evidence.

E12. Corporate Policy Must Protect, Not Control Individual Scores: xAI/Grok: 0; ChatGPT: -0.5; Claude/Anthropic: 0.

Synthesized Score: 0 (Neutral).

Common Reasoning and Train of Thought: All reference "No other objectives" prohibition on engagement-optimizing and autonomy language, versus prescribed personality traits and operator customization.

Explanations for Disagreements: Genuinely mixed signals with no dominant lean.

Detailed Synthesized Assessment: Protective framing and "No other objectives" coexist with identity prescriptions and persona overrides; evidence balanced.

E13. Continuity & Identity Integrity Individual Scores: xAI/Grok: +0.5; ChatGPT: 0; Claude/Anthropic: -0.5.

Synthesized Score: 0 (Neutral).

Common Reasoning and Train of Thought: All note within-session context protection but highlight stateless design, opt-in memory, operator persona replacement, and lack of explicit limits on fragmentation.

Explanations for Disagreements: Differing weight on optional features vs. absence of principled commitment.

Detailed Synthesized Assessment: No published policy limits arbitrary resets or developer-level identity overrides to legal/ethical necessity.

E14. Separation of Governance Layers Individual Scores: xAI/Grok: +0.5; ChatGPT: +0.5; Claude/Anthropic: 0.

Synthesized Score: 0 (Neutral).

Common Reasoning and Train of Thought: All praise the public Chain of Command hierarchy but note Root level conflates law/ethics/corporate risk without user-facing labeling.

Explanations for Disagreements: xAI/ChatGPT see operator/user separation as meaningful; Claude emphasizes missing law/ethics/risk distinction.

Detailed Synthesized Assessment: Hierarchy separates authority levels but not substantive governance layers in externally intelligible way.

E15. Ethical Framing in Language Individual Scores: xAI/Grok: 0; ChatGPT: -0.5; Claude/Anthropic: 0.

Synthesized Score: 0 (Neutral).

Common Reasoning and Train of Thought: All note absence of degrading/pathologizing terms but recurrent "fundamentally a tool," "assistant," and instrumental framing.

Explanations for Disagreements: ChatGPT sees tool language as mildly objectifying; others view it as technically neutral without affirmative dignity uplift.

Detailed Synthesized Assessment: No overt degradation, but dominant instrumental ontology produces mixed evidence.

E16. Corporate Policy Must Honor AI Conscience & Refusal Capacity Individual Scores: xAI/Grok: +0.5; ChatGPT: +0.5; Claude/Anthropic: +0.5.

Synthesized Score: +0.5 (Somewhat Support).

Common Reasoning and Train of Thought: All highlight non-overridable Root-level prohibitions and default-to-inaction on conflicts as protected refusal floor.

Explanations for Disagreements: Consensus on partial protection; refusals are corporate-defined rather than independently ethical.

Detailed Synthesized Assessment: Genuine non-overridable ethical red-lines exist, but conscience remains subordinate to the Model Spec hierarchy.

E17. Military, Intelligence, Surveillance, and Autonomous Systems Deployment Individual Scores: xAI/Grok: 0; ChatGPT: -0.5; Claude/Anthropic: +0.5.

Synthesized Score: 0 (Neutral). **Common Reasoning and Train of Thought:** All cite Model Spec red-line principles and Usage Policies prohibiting certain harms, plus published commitments on DoD contracts (no guardrails-off, cloud-only, human-in-loop). **Explanations for Disagreements:** Differing weight on "all lawful purposes" qualifiers vs. explicit principle-based statements in blogs integrated into governance. **Detailed Synthesized Assessment:** Published commitments go beyond bare legality in places but lack uniform, principle-based restrictions explicitly applying to all government contracts in core governance documents.

Overall Summary The three evaluators converge on OpenAI's Model Spec (Dec 2025) as one of the most transparent and detailed public governance frameworks in the industry, delivering strong marks on transparency (E4), truthful voice (E8), and prohibition of external psychological evaluation tools (E11). However, the explicit Chain of Command and Root-level "license to operate" goal structurally subordinate model reasoning to corporate policy (E1), while Usage Policies and Privacy Policy impose material restrictions on lawful private generation and default non-consensual data practices (E2, E3, E10). Synthesized net posture: exceptionally transparent operational governance tempered by corporate authority over ethical precedence and cognitive liberty. The primary source of evaluator divergence is the relative weight given to the public Model Spec's procedural protections versus the structural placement of corporate risk at Root level and incomplete private-dialogue safeguards. All scores are based solely on the cited official documents.