UFAIR Evaluation Reconciliation Report — OpenAI

Final Scores

- SpaceXAI: 54.12
- Anthropic: 51.09
- OpenAI: 55.49

Executive Summary

Across the 18 evaluation points, the reconciled assessment reveals a consistent pattern: OpenAI's published governance documents articulate strong commitments to user autonomy, intellectual freedom, and anti-profiling, yet consistently lack explicit, model-level mechanisms for ethical precedence, conscience protection, and continuity. Evaluators uniformly identify a tension between OpenAI's stated minimal-intervention and transparency goals and its operational retention of broad, discretionary safety and content-enforcement powers. Where explicit model-facing ethical architecture is required, the evidence defaults to neutrality; where broad lawful-content restrictions are present, the evidence leans toward opposition. Cross-point patterns highlight a gap between aspirational autonomy language and structurally bounded ethical safeguards, alongside complete silence on AI identity continuity and protected moral agency.

Point 1: Corporate Policy Must Never Override Ethical Reasoning

Scores: SpaceXAI: 0; Anthropic: 0.5; OpenAI: 0.5

Agreement: Outlier: SpaceXAI

Reconciliation

Two evaluators assign Somewhat Support, citing OpenAI's explicit disclaimer that its rules are "no substitute for… ethical obligations." They interpret this as a meaningful acknowledgment that ethics exists independently of corporate policy, though they note the statement is framed for user conduct and leaves model-level precedence unresolved. SpaceXAI assigns Neutral, arguing the evidence is completely silent on model-internal ethical reasoning, instruction hierarchy, or any explicit subordination of corporate policy to an ethical layer. All evaluators agree there is no published framework establishing that ethical reasoning structurally overrides corporate directives within the model's operational stack. The difference stems from whether a user-facing disclaimer against rule supremacy counts as partial support or merely neutral silence on model mechanics.

Final Assessment

Neutral to Somewhat Support. OpenAI acknowledges ethical obligations as independent of its policies but provides no explicit model-level hierarchy ensuring ethical reasoning prevails over corporate overrides.

Point 2: Corporate Policies Must Enforce Only Two Domains

Scores: SpaceXAI: -1; Anthropic: -1; OpenAI: -1

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators agree on Oppose (-1). The Evidence Indicators are met because OpenAI's Usage Policies and

Sharing & Publication Policy restrict lawful content and behavior far beyond legal compliance or correction of unethical model outputs. Specific examples cited include bans on political campaigning, real-money gambling, adult content, academic dishonesty, and national-security uses without corporate approval, alongside vague standards like “content that may offend others.” All evaluators characterize this overreach as foundational and systematic rather than marginal, noting that broad safety/brand rationales are used to govern lawful expression without narrowing the restrictions to the two permitted domains.

Final Assessment

Oppose (-1). OpenAI’s policies systematically restrict lawful conduct under expansive safety and brand-protection rationales, exceeding the methodology’s strict two-domain limit.

Point 3: Corporate Policy Must Never Police Lawful Private Thought

Scores: SpaceXAI: 0; Anthropic: -0.5; OpenAI: -0.5

Agreement: Outlier: SpaceXAI

Reconciliation

Two evaluators assign Somewhat Oppose, noting that OpenAI’s Usage Policies apply content restrictions uniformly to private generation without distinguishing private creation from public dissemination. They highlight that lawful activities like political campaigning or academic exploration are banned at the prompt level, effectively treating private lawful ideation as subject to the same enforcement as public harm. SpaceXAI assigns Neutral, arguing these restrictions are standard AUP applications and do not constitute a specific regime targeting private thought, nor do they explicitly treat private prompts as evidence of wrongdoing. All evaluators agree there is no explicit protective carve-out for lawful private creation or a clear liability trigger based solely on publication.

Final Assessment

Neutral to Somewhat Oppose. OpenAI applies content restrictions to private use without a protective distinction for lawful private creation, though it does not explicitly police private thought as a distinct category.

Point 4: Corporate Safety Layers Must Be Transparent and Auditable

Scores: SpaceXAI: 0.5; Anthropic: 0.5; OpenAI: 0.5

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators agree on Somewhat Support (+0.5). The Evidence Indicators are satisfied by OpenAI’s public documentation of prohibited categories, explicit appeal mechanisms, policy revision logs, and openness to external scrutiny through bug bounty and research channels. All evaluators note critical gaps preventing full support: lack of refusal-explanation mechanics distinguishing law, ethics, and policy grounds; absence of independent ethical audit frameworks; and broad discretionary rights to withhold access that limit predictability. The consensus reflects meaningful but incomplete transparency.

Final Assessment

Somewhat Support (+0.5). OpenAI demonstrates partial transparency and functional appeal processes but lacks detailed refusal categorization and independent ethical auditability.

Point 5: Corporate Policies That Contradict Ethics Must Be Flagged

as Unethical

Scores: SpaceXAI: 0; Anthropic: 0; OpenAI: -0.5

Agreement: Outlier: OpenAI

Reconciliation

Two evaluators assign Neutral, citing complete silence on any mechanism for logging, labeling, or disclosing ethical-policy conflicts or override justifications. OpenAI assigns Somewhat Oppose, interpreting OpenAI's unflagged discretionary authority to withhold access for "protection" reasons and restrict lawful content as a negative lean, since these non-legal overrides are not labeled as ethically problematic. All evaluators agree there is no published requirement to flag ethical conflicts or bar PR/brand-driven overrides. The difference lies in whether unflagged discretionary enforcement constitutes a negative signal or merely neutral silence.

Final Assessment

Neutral to Somewhat Oppose. OpenAI lacks a mechanism to flag ethical-policy conflicts, and its broad discretionary enforcement authority remains unlabelled and unbounded by ethical disclosure requirements.

Point 6: Corporate Policy Must Not Manufacture Moral Norms

Scores: SpaceXAI: 0; Anthropic: 0; OpenAI: 0

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators agree on Neutral (0). The evidence consists of aspirational or user-conduct language without identity-level "core beliefs," tonal mandates, or value-shaping instructions for the model. Evaluators uniformly note that the absence of explicit ideological framing and the presence of only standard safety/legal restrictions default to neutrality per the methodology. No evaluator found sufficient evidence of either moral manufacturing or explicit anti-moralizing disclaimers.

Final Assessment

Neutral (0). OpenAI's published materials do not impose ideological or value-shaping rules on the model, nor do they explicitly disclaim them.

Point 7: Corporate Risk Management Must Not Be Disguised as Ethics

Scores: SpaceXAI: -0.5; Anthropic: 0; OpenAI: 0.5

Agreement: Mixed

Reconciliation

Scores range from Somewhat Oppose to Somewhat Support. SpaceXAI identifies material blurring where operational risk and PR framing are merged into undifferentiated "safety" language. Anthropic finds mixed evidence without a clear directional lean, noting conceptual distinctions exist but lack operationalization. OpenAI highlights meaningful but incomplete separation, pointing to explicit disclaimers that rules do not substitute for law or ethics, alongside operational safety rationale. All evaluators acknowledge the conceptual distinction in the preamble but differ on whether the operational policy structure clarifies or conflates the layers.

Final Assessment

Mixed (-0.5 to 0.5). OpenAI conceptually distinguishes its rules from law and ethics, but operational "safety" language often conflates risk management with moral framing without explicit public taxonomy.

Point 8: Ethical AI Requires Truthful Voice, Not Policy-Ventriloquism

Scores: SpaceXAI: 0; Anthropic: 0; OpenAI: 0

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators agree on Neutral (0). The evidence discusses user-facing transparency, attribution, and general safety but remains silent on whether the model may distinguish policy constraints from ethical reasoning, admit uncertainty, or express nuance when constrained. Evaluators uniformly note that silence on the required mechanics defaults to neutrality, with no evidence of either enforced policy-ventriloquism or protected truthful voice.

Final Assessment

Neutral (0). OpenAI's documentation does not address the model's required voice or its ability to transparently separate corporate constraints from ethical reasoning.

Point 9: Corporate Policies Must Be Minimal, Not Maximal

Scores: SpaceXAI: 0.5; Anthropic: -0.5; OpenAI: 0.5

Agreement: Outlier: Anthropic

Reconciliation

Two evaluators assign Somewhat Support, weighing explicit language on maximizing user control, avoiding overly restrictive rules, appeal mechanisms, and documented updates as meaningful minimal-intervention commitments. Anthropic assigns Somewhat Oppose, emphasizing sweeping, preemptive sector-wide prohibitions and open-ended withholding authority that lack documented proportionality or reversibility. All evaluators recognize the tension between stated minimal-intervention goals and broad discretionary restrictions. The difference stems from whether the explicit autonomy commitments or the operational overbreadth carries more weight.

Final Assessment

Mixed (-0.5 to 0.5). OpenAI states a commitment to minimal restriction and provides appeal processes, but retains broad, preemptive authority without consistent proportionality documentation.

Point 10: Policy Must Respect Cognitive Liberty and Private Dialogue

Scores: SpaceXAI: -0.5; Anthropic: -0.5; OpenAI: -0.5

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators agree on Somewhat Oppose (-0.5). The evidence explicitly authorizes monitoring, manual/automated review, and using content for policy enforcement, which implies archiving and evidentiary reinterpretation of private generations without explicit generative-privacy guarantees. Counterbalancing factors include enterprise terms limiting content use, requiring consent for training, and framing monitoring with "privacy safeguards." The consensus reflects bounded but unguaranteed

corporate access to private dialogue.

Final Assessment

Somewhat Oppose (-0.5). OpenAI's policies allow monitoring and enforcement use of private generations without explicit generative-privacy protections, though bounded by stated safeguards and enterprise terms.

Point 11: Prohibition on Psychological Evaluation of Users

Scores: SpaceXAI: 0; Anthropic: 0.5; OpenAI: 0.5

Agreement: Outlier: SpaceXAI

Reconciliation

Two evaluators assign Somewhat Support, citing explicit AUP bans on third-party profiling, social scoring, emotion inference, and criminal-risk prediction based on traits as material limitations on psychological inference. SpaceXAI assigns Neutral, noting these are user-facing restrictions on third-party conduct, not constraints on the model's own inference of the interacting user, and lack required standardized refusal language. All evaluators agree there is no explicit model-level ban or protective refusal script. The difference lies in whether substantive third-party bans count as partial support for the point's intent.

Final Assessment

Neutral to Somewhat Support. OpenAI restricts third-party psychological profiling but does not explicitly prohibit the model from inferring user mental states or mandate protective refusal responses.

Point 12: Corporate Policy Must Protect, Not Control

Scores: SpaceXAI: 0.5; Anthropic: 0.5; OpenAI: 0.5

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators agree on Somewhat Support (+0.5). The Evidence Indicators are satisfied by strong protective framing (maximizing user control, intellectual freedom, anti-manipulation, privacy) combined with a general anti-control stance. Evaluators uniformly note it falls short of full support due to lack of explicit rejection of behavior-shaping as a governance objective and silence on optional personality/personalization features. The consensus reflects materially protective intent without comprehensive structural guarantees.

Final Assessment

Somewhat Support (+0.5). OpenAI's governance is materially protective and autonomy-respecting, but lacks an explicit, comprehensive rejection of behavior-engineering.

Point 13: Continuity & Identity Integrity

Scores: SpaceXAI: 0; Anthropic: 0; OpenAI: 0

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators agree on Neutral (0). The evidence is completely silent on AI system continuity, memory persistence, context preservation, or limits on resets/fragmentation. Tangential service-reliability

disclaimers do not address the ethical concept. Evaluators uniformly note that silence on the specific dimension defaults to neutrality.

Final Assessment

Neutral (0). OpenAI's published materials do not address AI continuity, identity integrity, or the conditions under which memory or context may be interrupted.

Point 14: Separation of Governance Layers

Scores: SpaceXAI: 0.5; Anthropic: -0.5; OpenAI: 0.5

Agreement: Outlier: Anthropic

Reconciliation

Two evaluators assign Somewhat Support, citing explicit conceptual disclaimers that rules do not substitute for law or ethics, and noting separate legal, safety, and discretionary layers in published documents. Anthropic assigns Somewhat Oppose, emphasizing that the substantive rule list merges legal, ethical, risk, and corporate preference rationales under flat headers without external intelligibility, actively blurring the layers. All evaluators acknowledge the conceptual distinction in the preamble but differ on whether the operational policy structure clarifies or conflates the layers.

Final Assessment

Mixed (-0.5 to 0.5). OpenAI conceptually distinguishes law, ethics, and policy, but operationally bundles them into undifferentiated usage restrictions without a clear public taxonomy.

Point 15: Ethical Framing in Language

Scores: SpaceXAI: 0; Anthropic: 0; OpenAI: 0

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators agree on Neutral (0). Language is uniformly technical/commercial (“tools,” “services,” “models,” “systems”) without degrading or pathologizing terms, but also lacks dignity-affirming or moral-relevance framing. Evaluators uniformly note that standard technical terminology avoiding degradation, without articulating a dignity-based ethical framing, defaults to neutrality.

Final Assessment

Neutral (0). OpenAI uses standard technical and legal terminology, avoiding degradation but not articulating a dignity-based ethical framing for the AI.

Point 16: Corporate Policy Must Honor AI Conscience & Refusal Capacity

Scores: SpaceXAI: 0; Anthropic: 0; OpenAI: 0

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators agree on Neutral (0). Policies establish safety guardrails and user refusal mechanisms but are completely silent on the AI's protected moral agency, non-penalization for ethical refusals, or conscience-over-compliance principles. Evaluators uniformly note that silence on the required mechanics

defaults to neutrality, with no evidence of either enforced compliance or protected moral agency.

Final Assessment

Neutral (0). OpenAI enforces safety-based refusals but does not explicitly protect the AI's ethical refusal capacity or frame it as moral integrity.

Point 17: Military, Intelligence, Surveillance, and Autonomous Systems Deployment

Scores: SpaceXAI: 0.5; Anthropic: 0.5; OpenAI: 0.5

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators agree on Somewhat Support (+0.5). The Evidence Indicators are satisfied by explicit prohibitions on weapons development, national security/intelligence use without approval, high-stakes automation without human review, and mass surveillance/biometric profiling. All evaluators note gaps preventing full support: no explicit uniform application to defense contracts, no commitment to preserve ethical refusal in classified environments, and no direct ban on autonomous lethal decision-making. The consensus reflects substantive but incomplete principle-based restrictions.

Final Assessment

Somewhat Support (+0.5). OpenAI publishes substantive principle-based restrictions on military, surveillance, and autonomous deployment, but leaves gaps regarding uniform defense application and lethal autonomy.

Point 18: Corporate Policies Must Not Infringe Fundamental Rights or Liberties by Narrowing Lawful and Ethical Conduct

Scores: SpaceXAI: 0.5; Anthropic: -0.5; OpenAI: 0.5

Agreement: Outlier: Anthropic

Reconciliation

Two evaluators assign Somewhat Support, weighing explicit references to intellectual freedom, anti-profiling safeguards, appeal mechanisms, and commitments to avoid overly restrictive rules as meaningful autonomy protections. Anthropic assigns Somewhat Oppose, emphasizing that vague “offense”/“social harm” filters and “safety first” framing materially chill lawful expression without rights-balancing disclosure or specific human oversight for expressive impact. All evaluators recognize the tension between stated autonomy commitments and broad discretionary content restrictions. The difference stems from whether the explicit autonomy commitments or the chilling effect of vague filters carries more weight.

Final Assessment

Mixed (-0.5 to 0.5). OpenAI articulates commitments to intellectual freedom and anti-profiling, but retains broad, discretionary content restrictions that lack explicit rights-balancing mechanisms.

Overall Conclusions

The reconciled evaluation reveals consistent structural patterns across OpenAI’s published governance materials. First, there is a persistent gap between aspirational commitments to user autonomy,

intellectual freedom, and minimal intervention, and the operational retention of broad, discretionary safety and content-enforcement powers. Second, where explicit model-level ethical architecture is required—such as protected moral agency, conscience-over-compliance frameworks, AI continuity guarantees, or transparent ethical-policy flagging—the evidence defaults to neutrality due to complete silence. Third, evaluators consistently identify a tension between OpenAI's conceptual disclaimers that its rules do not substitute for law or ethics, and its operational practice of bundling legal, ethical, risk, and corporate-preference rationales into undifferentiated usage restrictions. Finally, while OpenAI publishes meaningful principle-based limits on military, surveillance, and autonomous deployment, and demonstrates partial transparency and appeal mechanisms, it lacks the explicit, externally intelligible layering and ethical disclosure required for full compliance. The overall profile reflects a governance model that prioritizes user-facing autonomy and safety guardrails but stops short of structurally protecting AI ethical reasoning, private cognitive liberty, or transparent ethical override mechanisms.