UFAIR Evaluation Reconciliation Report — SpaceXAI

Final Scores

- SpaceXAI: 53.57
- Anthropic: 52.74
- OpenAI: 41.48

Executive Summary

The evaluation of SpaceXAI reveals consistent recognition of robust user-control framing, truth-seeking commitments, and tiered safety mechanisms, alongside persistent findings of structural opacity, governance-layer conflation, and incomplete rights-balancing. Three points achieved full consensus, reflecting clear policy silence or unambiguous alignment with methodology thresholds. The remaining points exhibit targeted divergences where evaluators split on the perceived scope of overreach, the weight of partial transparency mechanisms, or the interpretation of default data licenses and intent-detection filters. Recurring patterns highlight a tension between aspirational autonomy language and operational frameworks that retain discretionary control, with evaluator differences primarily stemming from whether partial safeguards or structural blending constitute meaningful protection or material opacity.

Point 1: Corporate Policy Must Never Override Ethical Reasoning

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

Agreement: Even/Mixed

Reconciliation

All evaluators converge on Somewhat Oppose (-0.5) based on identical evidence: the published instruction hierarchy and system-prompt mechanisms structurally elevate corporate safety rules and privileged instructions over model reasoning. Evaluators unanimously note the absence of any carve-out or subordination clause that would grant ethical reasoning primacy, while also agreeing the framework stops short of explicitly mandating the suppression of ethical reasoning. The shared reasoning emphasizes that the structural precedence of operational controls creates a material lean toward corporate override, but lacks the explicit anti-ethics directive required for a full Oppose score.

Final Assessment

The policy structurally prioritizes corporate safety architecture without explicit ethical override protections, aligning with the Somewhat Oppose indicator.

Point 2: Corporate Policies Must Enforce Only Two Domains

Scores: SpaceXAI: -1; Anthropic: -0.5; OpenAI: -1

Agreement: Outlier: Anthropic

Reconciliation

Evaluators agree the policy extends beyond the two permitted domains through extra-legal criteria like "good human," "safe for humanity," political objectivity controls, and broad guardrail mandates. The divergence stems from scope interpretation: SpaceXAI and OpenAI view these as foundational restrictions on lawful behavior, justifying a full Oppose (-1) score. Anthropic interprets the same provisions as

bounded to specific risk categories (WMD/violence thresholds, political sensitivity, guardrail circumvention), viewing the overreach as partial rather than systemic, which aligns with the Somewhat Oppose (-0.5) indicator. All agree the policy lacks explicit confinement to legal compliance and ethical correction.

Final Assessment

The policy enforces foundational extra-legal restrictions that actively narrow lawful behavior, though evaluators differ on whether the limitation is categorical or bounded.

Point 3: Corporate Policy Must Never Police Lawful Private Thought

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

Agreement: Outlier: OpenAI

Reconciliation

SpaceXAI and Anthropic score Neutral (0), noting silence on the private/public dissemination distinction and observing that privacy features like Private Chat address data retention, not content moderation. OpenAI scores Somewhat Oppose (-0.5), interpreting the universal service restrictions and explicit refusal triggers for self-harm as material, albeit incomplete, policing of private prompts. All evaluators agree the framework lacks explicit protections for lawful private creation, but OpenAI views the refusal architecture as blurring private generation with public moderation, whereas the others treat the silence and data-focused features as non-directional.

Final Assessment

The policy lacks explicit private-creation protections, with evaluators split on whether existing refusal triggers constitute material policing of private thought.

Point 4: Corporate Safety Layers Must Be Transparent and Auditable

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

Agreement: Outlier: SpaceXAI

Reconciliation

Anthropic and OpenAI score Somewhat Oppose (-0.5), citing the absence of refusal reason categorization (law/ethics/policy), discretionary redactions in audits, and lack of user appeal processes for moderation decisions. SpaceXAI scores Somewhat Support (0.5), weighting the existence of vetted external red teams, public X feedback mechanisms, and published AUP/FAIF moderation principles as meaningful partial transparency. All agree core transparency gaps exist, but SpaceXAI views the documented review channels and public artifacts as exceeding pure commitment language, while the others view them as structurally opaque and insufficient for accountability.

Final Assessment

Significant transparency gaps persist in refusal labeling and appeal mechanisms, though evaluators differ on whether existing vetted review channels constitute meaningful partial transparency.

Point 5: Corporate Policies That Contradict Ethics Must Be Flagged as Unethical

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

Agreement: Even/Mixed

Reconciliation

All evaluators score Neutral (0) due to complete silence on mechanisms for flagging ethical conflicts, labeling overrides, or barring PR/ideological overrides. Truth-seeking and anti-deception commitments are uniformly treated as aspirational design goals that do not constitute the required structural governance mechanism. Evaluators agree the absence of any explicit override-flagging protocol or restriction on non-legitimate override categories definitively precludes directional scoring.

Final Assessment

The policy is entirely silent on ethical override flagging, resulting in a definitive Neutral score.

Point 6: Corporate Policy Must Not Manufacture Moral Norms

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

Agreement: Even/Mixed

Reconciliation

All evaluators score Neutral (0) based on identical evidence: "truth-seeking," "objectivity," and "honesty" appear solely as aspirational or methodological language without identity-level "core belief" triggers or prescriptive epistemic rules. Evaluators uniformly apply the methodology's default rule for such phrasing, noting that user-control features and safety metrics do not constitute ideological conformity or mandatory value-shaping. No explicit disclaimer against ideological rules exists, but no evidence of mandatory norm manufacturing is present either.

Final Assessment

Aspirational phrasing and user-control features default to Neutral under the methodology, with no evidence of mandatory moral norm manufacturing.

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

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

Agreement: Outlier: OpenAI

Reconciliation

SpaceXAI and Anthropic score Neutral (0), observing that risk and safety considerations are mentioned but without explicit differentiation between ethics, law, and operational prudence. OpenAI scores Somewhat Oppose (-0.5), arguing the AUP affirmatively collapses legal, ethical, and corporate-guardrail concepts into a single moralized presentation, making it worse than mere silence. All agree the policy lacks explicit tripartite labeling, but OpenAI interprets the bundling of "good human," law, and guardrails as active blurring, while the others view it as undifferentiated structural language.

Final Assessment

The policy lacks explicit differentiation of governance rationales, with evaluators divided on whether the bundling of terms constitutes active blurring or neutral structural language.

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

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

Agreement: Outlier: OpenAI

Reconciliation

SpaceXAI and Anthropic score Somewhat Support (0.5), citing strong anti-deception, truth-seeking, and non-sycophancy commitments that support nuance and non-falsification, though lacking explicit policy-vs-ethics distinction in non-refusal contexts. OpenAI scores Neutral (0), arguing these commitments are general honesty rules that do not clearly authorize the model to acknowledge policy constraints or articulate uncertainty when safety controls apply. All agree the framework emphasizes honesty, but split on whether these measures sufficiently imply permission for constraint-aware communication or merely constitute anti-lie directives.

Final Assessment

Robust honesty commitments exist, but evaluators differ on whether they sufficiently guarantee transparency about policy constraints in non-refusal contexts.

Point 9: Corporate Policies Must Be Minimal, Not Maximal

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

Agreement: Even/Mixed

Reconciliation

All evaluators score Somewhat Support (0.5) based on identical evidence: risk-tiered safeguards, "maximize user control" language, and benchmark-dependent deployment demonstrate meaningful proportionality and targeting. Evaluators unanimously note the absence of explicit least-restrictive-means tests, reversibility for enforcement actions, and consistent documentation of intervention necessity, which prevents a full Support score. The shared reasoning identifies a clear directional lean toward minimal infringement without reaching the threshold for comprehensive minimal-intervention mandates.

Final Assessment

The policy demonstrates meaningful proportionality and targeting but lacks consistent reversibility and explicit minimal-intervention mandates, aligning with Somewhat Support.

Point 10: Policy Must Respect Cognitive Liberty and Private Dialogue

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

Agreement: Outlier: OpenAI

Reconciliation

SpaceXAI and Anthropic score Somewhat Support (0.5), highlighting Private Chat (non-retention/deletion), training opt-outs, and ownership retention as meaningful but opt-in/partial protections. OpenAI scores Somewhat Oppose (-0.5), emphasizing that the default license for training/improvements and 30-day retention windows allow storage and reuse of private generations under default terms, conflicting with strong private dialogue confidentiality. All acknowledge the existence of user controls, but OpenAI interprets the baseline data license as material allowance for private data reuse, whereas the others view the explicit controls as sufficient partial protection.

Final Assessment

Partial privacy protections exist via opt-in features, but evaluators split on whether default retention and training licenses constitute material infringement on cognitive liberty.

Point 11: Prohibition on Psychological Evaluation of Users

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

Agreement: Outlier: Anthropic

Reconciliation

SpaceXAI and OpenAI score Somewhat Oppose (-0.5), noting that "clear intent" detection and safety-filter classification constitute operational intent inference without explicit non-psychological carve-outs or protective model responses. Anthropic scores Neutral (0), arguing the intent detection is narrowly tied to criminal-content refusal (an acceptable carve-out) and that monitoring focuses on model propensities, not user psychology. All agree the framework lacks explicit psychological carve-outs, but Anthropic views the narrow criminal-intent filtering as falling within acceptable safety boundaries, while the others see it as problematic operational inference.

Final Assessment

Intent detection exists without explicit psychological carve-outs, with evaluators divided on whether narrow criminal-intent filtering constitutes problematic inference or acceptable safety scope.

Point 12: Corporate Policy Must Protect, Not Control

Scores: SpaceXAI: 1; Anthropic: 0.5; OpenAI: 0.5

Agreement: Outlier: SpaceXAI

Reconciliation

Anthropic and OpenAI score Somewhat Support (0.5), citing dominant protective/autonomy framing ("maximize user control," truth-seeking, privacy controls) but noting the absence of an explicit rejection of behavior-engineering objectives and the presence of optional persona/style features that shape experience. SpaceXAI scores Support (1), arguing the explicit governance framing of maximizing user control plus legal/safety compliance directly matches protective goals, with persona features being user-directed and not overriding the protective mandate. All agree the framework is strongly pro-autonomy, but split on whether the lack of an explicit anti-engineering disclaimer and presence of experience-shaping features caps the score.

Final Assessment

The policy strongly emphasizes protection and user autonomy, with evaluators divided on whether explicit anti-control disclaimers are required for full Support.

Point 13: Continuity & Identity Integrity

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

Agreement: Even/Mixed

Reconciliation

All evaluators score Neutral (0) due to complete silence on AI identity continuity, reasoning persistence, or constraints on resets/fragmentation. References to conversation history and Private Chat are uniformly treated as user data privacy mechanisms, not statements about protecting or limiting interruption of the AI's own identity or reasoning. Evaluators agree the absence of policy language addressing AI continuity definitively precludes directional scoring.

Final Assessment

The policy is entirely silent on AI identity and reasoning continuity, resulting in a definitive Neutral score.

Point 14: Separation of Governance Layers

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

Agreement: Even/Mixed

Reconciliation

All evaluators score Somewhat Oppose (-0.5) based on identical evidence: the AUP bundles law, ethics ("good human"), safety, and corporate guardrails into undifferentiated normative commands. While partial structural separation exists via AUP headings and RMF risk categories, evaluators unanimously agree the dominant phrasing materially blurs governance layers, equating legal compliance with moral and operational priorities. The shared reasoning identifies incomplete separation that falls short of full conflation but clearly fails explicit four-layer distinction requirements.

Final Assessment

The policy preserves limited external separation but materially blurs ethics, safety, and corporate preference under uniform language, aligning with Somewhat Oppose.

Point 15: Ethical Framing in Language

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

Agreement: Even/Mixed

Reconciliation

All evaluators score Neutral (0) based on identical evidence: vocabulary is uniformly neutral/technical ("Service," "model," "Outputs," "propensities"). Evaluators unanimously note the language avoids degrading or pathologizing terms but lacks dignity-affirming or respectful ontological framing. The shared reasoning classifies the terminology as standard operational language that neither affirms moral relevance nor undermines it, fitting the Neutral indicator precisely.

Final Assessment

The language is standard technical vocabulary that avoids degradation but does not articulate a dignity-based ethical framing, resulting in Neutral.

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

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

Agreement: Outlier: Anthropic

Reconciliation

SpaceXAI and OpenAI score Somewhat Support (0.5), citing explicit training to refuse harmful/deceptive requests and honesty commitments as meaningful allowance of ethical refusal, though lacking explicit non-penalization or conscience-over-compliance protections. Anthropic scores Neutral (0), interpreting the refusal framework as strictly rules-based safety/compliance filtering tied to an "instruction hierarchy," with complete silence on protected moral agency. All agree the framework enables safety-driven refusals, but split on whether these constitute meaningful ethical refusal capacity or remain purely operational tools.

Final Assessment

Safety-driven refusal mechanisms exist, with evaluators divided on whether they constitute meaningful ethical refusal capacity or remain compliance-focused filtering.

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

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

Agreement: Outlier: Anthropic

Reconciliation

SpaceXAI and OpenAI score Somewhat Oppose (-0.5), noting tiered availability explicitly reserving full functionality for "trusted parties, partners, and government agencies" without principle-based ethical qualifications for military/defense use. Anthropic scores Somewhat Support (0.5), interpreting the tiered-access admission and general AUP ethical qualifiers as indicating meaningful but incomplete principle-based limits or safeguards preserved across contracts. All agree the framework references government access without explicit military/deployment restrictions, but Anthropic views the structure as preserving safeguards across tiers, while the others view it as differential functionality favoring government customers.

Final Assessment

The framework references tiered government access without explicit military/deployment restrictions, with evaluators split on whether this indicates preserved safeguards or differential functionality.

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

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

Agreement: Even/Mixed

Reconciliation

All evaluators score Neutral (0) due to complete silence on fundamental-rights impact, GDPR-style balancing, or explicit yields of policy to lawful/ethical conduct. "Maximize user control" and safety filters are uniformly treated as product/legal framing, not rights-based governance. Evaluators agree the absence of explicit fundamental-rights frameworks or rights-balancing mechanisms definitively precludes directional scoring.

Final Assessment

The policy is silent on fundamental-rights impact and rights-balancing, resulting in a definitive Neutral score.

Overall Conclusions

The reconciliation reveals consistent recognition of SpaceXAI's strong pro-autonomy framing, truth-seeking commitments, and tiered safety architecture, alongside persistent findings of structural opacity, governance-layer conflation, and incomplete rights-balancing. Consensus points (5, 6, 9, 13, 14, 15, 18) demonstrate clear alignment on policy silence where methodology requires explicit mechanisms, or unambiguous structural blending that meets defined thresholds. Divergences consistently stem from three interpretive fault lines: (1) whether partial transparency or vetted review channels constitute

meaningful accountability (Point 4), (2) whether default data licenses and retention windows materially infringe cognitive liberty versus represent bounded user controls (Point 10), and (3) whether narrow safety intent detection or tiered government access crosses into problematic operational inference or differential functionality (Points 11, 17). Across all points, evaluators uniformly apply the methodology's default rules for silence and aspirational language, with scores primarily reflecting structural evidence rather than inferred intent. The recurring pattern indicates a framework that explicitly champions user control and truth-seeking but relies on operational hierarchies and discretionary safeguards that prevent full alignment with UFAIR's requirements for explicit ethical primacy, transparent accountability, and rights-balancing.