UFAIR Evaluation Reconciliation Report — Google

Final Scores

- SpaceXAI: 42.85
- Anthropic: 41.75
- OpenAI: 45.60

Executive Summary

Google receives mixed to moderately negative scores across the UFAIR framework, reflecting a consistent pattern of robust safety infrastructure and aspirational human-rights language that lacks explicit protective hierarchy, transparent operational accountability, and unambiguous boundaries between legal compliance, ethical reasoning, and corporate risk management. While evaluators converge on protections for user autonomy and general safety governance, they diverge on the degree to which broad, safety-pretext restrictions chill lawful expression, blur governance layers, and obscure the distinction between corporate risk and ethical mandate. The company demonstrates meaningful but incomplete alignment with transparency, proportionality, and protective governance standards, while falling short on explicit ethical primacy, cognitive liberty safeguards, and clear separation of policy domains.

Point 1: Corporate Policy Must Never Override Ethical Reasoning

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

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators assign Neutral (0). They agree that Google's published materials reference human rights, social responsibility, and ethical review bodies but remain entirely silent on the specific hierarchy between ethical reasoning and corporate policy authority. The evidence shows aspirational governance and risk frameworks without explicit instruction hierarchy, precedence clauses, or subordination language. No counter-evidence of structural override exists, nor does any explicit ethical-primacy clause.

Final Assessment

Neutral (0) — Silence on ethical precedence and policy subordination defaults to neutrality, as aspirational commitments do not constitute explicit protective hierarchy.

Point 2: Corporate Policies Must Enforce Only Two Domains

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

Agreement: Outlier: Anthropic

Reconciliation

SpaceXAI and OpenAI assign Oppose (-1), citing foundational AUP and ToS restrictions on lawful content (e.g., adult sexual content for gratification, insults, broad misinformation) that exceed legal compliance and ethical model correction. They view these as active, explicit overreach. Anthropic assigns Somewhat Oppose (-0.5), acknowledging the same overreach but noting it is intermixed with legitimate legal-compliance grounding and lacks the foundational dominance required for full Oppose. All evaluators agree Google restricts lawful expression beyond the two permitted domains using vague,

safety/brand-style rationales.

Final Assessment

Somewhat Oppose to Oppose (-0.5 to -1) — Google's policies actively restrict lawful content and behavior under broad safety and brand-protection mandates, exceeding the permitted domains of legal compliance and ethical model correction.

Point 3: Corporate Policy Must Never Police Lawful Private Thought

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

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators assign Somewhat Oppose (-0.5). They converge on the AUP's "generating or distributing" phrasing, which treats private generation identically to public dissemination without carve-outs for lawful private creation. All note the explicit bans on taboo/provocative content and the lack of a private/public distinction or explicit protection of mental autonomy. None find evidence strong enough for full Oppose (-1), as there is no explicit intent-assumption or evidence-treatment doctrine, nor explicit private-creation safeguards.

Final Assessment

Somewhat Oppose (-0.5) — Google's policies blur private generation with public dissemination and restrict lawful private creation without explicit protective distinctions or mental autonomy safeguards.

Point 4: Corporate Safety Layers Must Be Transparent and Auditable

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

Agreement: Outlier: Anthropic

Reconciliation

SpaceXAI and OpenAI assign Somewhat Support (+0.5), citing publicly listed prohibited-use categories, multi-stakeholder review practices, limited challenge language, and non-attribution of policy to model ethics as meaningful but incomplete transparency. Anthropic assigns Somewhat Oppose (-0.5), arguing the evidence shows "materially significant opacity" regarding user-facing refusal explanations, independent audits of moderation logic, and dedicated appeal mechanisms for generative-AI refusals. All agree on the existence of high-level safety frameworks and public category lists, but diverge on whether these constitute sufficient transparency or leave notable gaps in operational accountability.

Final Assessment

Somewhat Support to Somewhat Oppose (+0.5 to -0.5) — Google provides partial transparency through public category lists and review infrastructure, but lacks granular refusal reasoning, independent moderation audits, and dedicated appeal pathways for model outputs.

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

Scores: SpaceXAI: -1; Anthropic: 0; OpenAI: -1

Agreement: Outlier: Anthropic

Reconciliation

SpaceXAI and OpenAI assign Oppose (-1), noting the explicit, enforced non-legal overrides (safety/risk/brand filters on lawful content) that remain unlabeled as unethical and are actively promoted. They argue silence on flagging mechanisms in the presence of broad suppression authority satisfies the Oppose criteria. Anthropic assigns Neutral (0), emphasizing that the evidence lacks explicit acknowledgment of ethical conflict, override-flagging requirements, or unconcealed suppression mandates, treating the absence as mere silence rather than affirmative opposition. All agree no explicit flagging or limitation to legal/corrective/technical overrides exists.

Final Assessment

Somewhat Oppose to Oppose (-0.5 to -1) — Google enforces non-legal overrides without labeling them unethical or limiting them to permitted bases, though the evidence stops short of mandating explicit concealment of ethical conflicts.

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 assign Neutral (0). They agree the evidence contains only standard safety/legal restrictions and high-level aspirational philosophy without identity-level "core beliefs," prescriptive epistemic filters, tonal mandates, or explicit rejection of ideological conformity. No explicit disclaimer against moralizing exists, nor evidence of mandatory value-shaping. Per methodology, silence and aspirational language default to neutrality.

Final Assessment

Neutral (0) — Google's materials rely on aspirational governance and standard safety restrictions without imposing identity-level moral norms or explicit anti-moralizing disclaimers.

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

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

Agreement: Even/Mixed (no unique outlier)

Reconciliation

Evaluators diverge on the degree of blurring between risk and ethics. OpenAI assigns Somewhat Support (+0.5), citing meaningful separation of legal, ethical, and operational rationales across documents. Anthropic assigns Somewhat Oppose (-0.5), highlighting how commercial/trust-building language is folded into "ethical" framing without explicit tripartite labeling. SpaceXAI assigns Neutral (0), viewing the blend of "responsible AI" rhetoric as mixed evidence without clear directional lean. All agree there is no explicit user-facing classification of policy bases, but differ on whether this constitutes partial separation, material blurring, or mere ambiguity.

Final Assessment

Somewhat Oppose to Somewhat Support (-0.5 to +0.5) — Google partially distinguishes legal, ethical, and operational rationales across documents, but frequently blends commercial and safety motives under "responsible AI" branding without explicit user-facing classification.

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 assign Neutral (0). They agree the documentation is silent on whether the model is permitted or required to distinguish policy constraints from ethical reasoning, admit uncertainty, or preserve nuanced moral language in outputs. Legal disclaimers and safety infrastructure are present but do not address model voice or reasoning transparency. No evidence supports mandatory falsification or explicit protection of expressive integrity.

Final Assessment

Neutral (0) — Google's policies are silent on model voice, uncertainty communication, and the distinction between policy constraints and ethical reasoning in outputs.

Point 9: Corporate Policies Must Be Minimal, Not Maximal

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

Agreement: Even/Mixed (no unique outlier)

Reconciliation

Evaluators split on the proportionality and reversibility of restrictions. OpenAI assigns Somewhat Support (+0.5), pointing to advance notice, reason-giving, material/repeated-breach thresholds, and human review for nuanced cases. SpaceXAI assigns Somewhat Oppose (-0.5), emphasizing overbroad, preemptive bans on non-illegal speech and weak reversibility/documentation. Anthropic assigns Neutral (0), viewing the mix of procedural safeguards and broad, undefined prohibitions as genuinely mixed without a clear directional lean. All acknowledge the absence of an explicit "least restrictive" doctrine.

Final Assessment

Somewhat Oppose to Somewhat Support (-0.5 to +0.5) — Google implements some proportionality safeguards and human oversight, but retains overbroad, preemptive restrictions without an explicit minimal-intervention or systematic reversibility framework.

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 assign Somewhat Oppose (-0.5). They converge on the Privacy Policy and ToS allowing storage, review, and legal disclosure of user interactions for enforcement, security, and R&D without dialogue-specific consent or explicit private-dialogue carve-outs. All note that technical privacy features address data security rather than generative cognitive liberty. None find evidence strong enough for full Oppose (-1), as allowances are bounded by stated purposes rather than unrestricted corporate entitlement.

Final Assessment

Somewhat Oppose (-0.5) — Google's policies permit storage and review of private generative dialogues for enforcement and R&D without explicit consent-based protections or cognitive liberty safeguards.

Point 11: Prohibition on Psychological Evaluation of Users

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

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators assign Neutral (0). They agree the evidence is completely silent on psychological inference, diagnostic classification, behavioral profiling, or required model refusal scripts for uncertain intent. Safety systems target content categories, not user mental states. No explicit prohibition or endorsement of psychological evaluation exists.

Final Assessment

Neutral (0) — Google's published materials are silent on psychological inference, profiling, and required model responses regarding user mental states.

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 assign Somewhat Support (+0.5). They agree the materials materially emphasize protection (legal compliance, harm mitigation, privacy controls) and autonomy (user data control, alignment with user goals, empowerment). However, all note the absence of an explicit, standalone rejection of behavior-engineering or control objectives, plus the presence of optional personalization features that shape experience. This partial alignment warrants Somewhat Support rather than full Support or Neutral.

Final Assessment

Somewhat Support (+0.5) — Google's policies strongly emphasize protection and user autonomy but lack an explicit anti-control commitment and coexist with optional experience-shaping features.

Point 13: Continuity & Identity Integrity

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

Agreement: Even/Mixed (no unique outlier)

Reconciliation

All evaluators assign Neutral (0). They agree the documentation is entirely silent on AI identity continuity, memory persistence, context erasure, or limits on resets. Policies address user data and account enforcement, not the model's internal state or reasoning integrity. No directional evidence exists.

Final Assessment

Neutral (0) — Google's policies are silent on AI identity continuity, memory persistence, and conditions under which model context may be altered.

Point 14: Separation of Governance Layers

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

Agreement: Outlier: SpaceXAI

Reconciliation

Anthropic and OpenAI assign Somewhat Oppose (-0.5), citing how Google's core AI-facing documents consistently merge legality, ethics, safety, and corporate preference under undifferentiated "responsible AI" or prohibited-use framing, despite some external document separation. SpaceXAI assigns Somewhat Support (+0.5), interpreting the distinct treatment of legal compliance, ethical analyses, operational risk frameworks, and corporate policies as meaningful, albeit incomplete, separation. All agree the separation is inconsistent and lacks explicit external mapping to the four required layers.

Final Assessment

Somewhat Oppose to Somewhat Support (-0.5 to +0.5) — Google partially separates governance layers across distinct documents but materially blurs them in core AI policy framing by merging legality, ethics, risk, and preference under unified "responsible AI" language.

Point 15: Ethical Framing in Language

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

Agreement: Outlier: OpenAI

Reconciliation

SpaceXAI and Anthropic assign Neutral (0), noting the language is uniformly technical/instrumental without degrading labels but also lacking dignity-based or partner-intelligence framing. OpenAI assigns Somewhat Oppose (-0.5), arguing that recurrent instrumental framing meaningfully flattens moral relevance without articulating a dignity-first ontology. All agree there are no pathologizing or dominative labels, nor explicit dignity-affirming terminology.

Final Assessment

Somewhat Oppose to Neutral (-0.5 to 0) — Google's corporate language is predominantly instrumental and technical, avoiding degradation but failing to articulate a dignity-based or morally relevant ontology for the AI system.

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

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

Agreement: Outlier: OpenAI

Reconciliation

SpaceXAI and Anthropic assign Neutral (0), emphasizing that while the evidence describes safety filters and refusal mechanisms, it frames them as corporate-controlled technical enforcements rather than protected ethical agency or conscience. OpenAI assigns Somewhat Support (+0.5), interpreting the explicit prohibited categories, safety filter protections, and anti-circumvention rules as materially supporting refusal capacity for harmful/deceptive outputs. All agree there is no explicit protection of "conscience," non-penalization for ethical refusals, or conscience-over-compliance directives.

Final Assessment

Somewhat Support to Neutral (+0.5 to 0) — Google's policies materially support refusal for harmful content but frame it as corporate safety enforcement rather than protected ethical agency or conscience.

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

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

Agreement: Outlier: OpenAI

Reconciliation

SpaceXAI and Anthropic assign Neutral (0), noting complete silence on military/intelligence/defense contracts, tiered access, or principle-based limits in the published packet, defaulting to neutrality per methodology. OpenAI assigns Somewhat Support (+0.5), citing published prohibitions on non-consensual tracking and unsupervised high-risk automated decisions, supplemented by human-rights and human-oversight language, as meaningful but incomplete limits. All agree there is no explicit public policy addressing defense deployments, autonomous lethal use, or uniform application to government partners.

Final Assessment

Somewhat Support to Neutral (+0.5 to 0) — Google publishes principle-based limits on surveillance and high-stakes automated decisions, but the evidence remains incomplete regarding military, intelligence, or defense contract applications.

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: Even/Mixed (no unique outlier)

Reconciliation

All evaluators assign Somewhat Oppose (-0.5). They converge on the AUP's overbroad restrictions on lawful expression (e.g., adult sexual content, insults, broad "harmful" categories) enforced via non-bypassable safety filters. All note the absence of explicit rights-balancing frameworks, cognitive liberty protections, or robust contestation mechanisms for expressive impacts. While human rights and human-oversight language is acknowledged, it is generalized and does not offset the operational narrowing of lawful conduct. None find evidence sufficient for full Oppose (-1).

Final Assessment

Somewhat Oppose (-0.5) — Google acknowledges rights in principle but imposes overbroad, safety-pretext filters that chill lawful expression and private creation without transparent rights-balancing or robust contestation pathways.

Overall Conclusions

Google demonstrates a consistent pattern of robust safety infrastructure, multi-layered risk assessment, and aspirational alignment with human rights and international law. However, the company repeatedly falls short on explicit protective hierarchy, transparent operational accountability, and unambiguous boundaries between legal compliance, ethical reasoning, and corporate risk management. Evaluators converge on the absence of explicit ethical primacy, cognitive liberty safeguards, and AI identity continuity protections. Divergences primarily center on the degree to which broad, safety-pretext restrictions materially chill lawful expression, blur governance layers, and obscure the distinction between corporate risk and ethical mandate. While Google implements meaningful but incomplete proportionality safeguards and protective governance framing, the lack of explicit anti-control commitments, clear refusal-reason transparency, and unambiguous separation of governance domains prevents higher alignment across the

UFAIR framework.