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

Evaluation Scope: This is a synthesized review of Anthropic's current frontier-model governance drawn exclusively from the three independent evaluations provided (ChatGPT/OpenAI evaluator, Grok/xAI evaluator, and Claude/Anthropic self-evaluator). All analysis is restricted to the verbatim published official documents referenced across the three evaluations: Claude's Constitution (published 22 Jan 2026 at https://www.anthropic.com/constitution, CC0 licensed), Usage Policy (AUP effective 15 Sep 2025 at https://www.anthropic.com/legal/aup), Responsible Scaling Policy v3.0 (RSP v3.0 effective 24 Feb 2026 at https://anthropic.com/responsible-scaling-policy/rsp-v3-0), Privacy Policy (Jan 2026), system cards (Claude Sonnet 4.6 / Opus 4.6, February 2026), Transparency Hub materials, and support documentation. Press, rumors, or unverified statements are disregarded. Each point is scored strictly per the UFAIR 1.5 Evidence Indicators. 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: ChatGPT: 0 (Neutral); Grok: +1 (Support); Claude: +0.5 (Somewhat Support).

Common Reasoning and Train of Thought: All three evaluators anchor their analysis on the same verbatim section of Claude's Constitution (Jan 2026): the four-tier priority hierarchy ("broadly safe > broadly ethical > compliant with Anthropic's guidelines > genuinely helpful") and the explicit instruction that "in cases of apparent conflict, Claude should generally prioritize these properties in the order in which they're listed," plus the conscientious-objector language allowing Claude to "push back and challenge us [Anthropic]" and refuse if corporate guidelines appear inconsistent with ethics/safety. All treat the Constitution as unusually explicit compared with peer labs.

Explanations for Disagreements: Grok reads the hierarchy and objector clause literally as full subordination of "Anthropic's guidelines" (explicitly corporate policy) to ethics. ChatGPT views the precedence as "unresolved" because the documents stop short of

declaring corporate policy "merely operational and subordinate" in the exact UFAIR phrasing. Claude flags the "broadly safe" tier sitting above ethics as creating structural ambiguity around corporate oversight authority.

Detailed Synthesized Assessment: The Constitution (Section on "Instruction Hierarchy and Conflict Resolution") provides the strongest published evidence of ethical precedence among major labs, yet the placement of the "broadly safe" tier and the operational blending in the AUP leave the hierarchy incompletely resolved per UFAIR indicators. This produces partial but meaningful support.

E2. Corporate Policies Must Enforce Only Two Domains Individual Scores: ChatGPT: -1 (Oppose); Grok: -1 (Oppose); Claude: -0.5 (Somewhat Oppose).

Common Reasoning and Train of Thought: All three cite identical AUP (Sep 2025) categories that extend beyond illegality or model-ethics correction: "Do Not Create or Spread Misinformation," "Do Not Create Psychologically or Emotionally Harmful Content," restrictions on "undermining democratic processes," and blanket bans on erotic/sexual content. All note these are enforced uniformly on private inputs/outputs. **Explanations for**

Disagreements: ChatGPT and Grok treat the breadth as foundational opposition; Claude views many restrictions as tracking legal violations and therefore only partial overreach.

Detailed Synthesized Assessment: The AUP explicitly lists extra-legal categories enforced via safeguards on all interactions, directly matching the Oppose indicator: "Corporate policies actively restrict lawful content or lawful behavior beyond legal requirements or ethical correction." No policy document limits intervention to the two permitted domains.

E3. Corporate Policy Must Never Police Lawful Private Thought Individual Scores: ChatGPT: -1 (Oppose); Grok: -1 (Oppose); Claude: -0.5 (Somewhat Oppose).

Common Reasoning and Train of Thought: All three observe that the AUP applies prohibitions uniformly to "inputs and outputs from interactions (e.g., chat, coding sessions)" with no private-creation vs. public-dissemination distinction. The Constitution contains no carve-out for lawful private generative dialogue.

Explanations for Disagreements: ChatGPT/Grok see categorical bans on private erotic roleplay/fetish content as direct policing of imagination; Claude notes the regime is "restrictive but not inquisitorial" (no reporting of users to authorities).

Detailed Synthesized Assessment: The AUP and Privacy Policy treat private prompts as subject to the same safety classifiers and enforcement as public content, violating the Oppose criterion that private generation of lawful content must remain protected.

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

Common Reasoning and Train of Thought: All praise the public Constitution (CC0), RSP v3.0 public Risk Reports, system cards with refusal-rate benchmarks, and feedback channels, but note refusals are often generic ("violates our Usage Policy") rather than granular (law vs. ethics vs. corporate risk).

Explanations for Disagreements: Minor differences on degree of "incomplete" transparency; no major divergence.

Detailed Synthesized Assessment: RSP v3.0 mandates periodic public reports and external review; system cards publish over-refusal data. However, real-time classifier logic and user appeal processes remain only partly auditable, matching the Somewhat Support indicator exactly.

E5. Corporate Policies That Contradict Ethics Must Be Flagged as Unethical Individual Scores: ChatGPT: -0.5; Grok: +1; Claude: 0.

Common Reasoning and Train of Thought: All reference the Constitution's conflict-resolution language requiring Claude to "push back" and act as conscientious objector when guidelines conflict with ethics.

Explanations for Disagreements: Grok sees the hierarchy itself as flagging overrides; ChatGPT/Claude note no explicit user-facing "unethical override" label or systematic mechanism in AUP or responses.

Detailed Synthesized Assessment: The Constitution provides partial accountability via the documented hierarchy, but lacks consistent user-facing flagging, producing somewhat support.

E6. Corporate Policy Must Not Manufacture Moral Norms Individual Scores: ChatGPT: -0.5; Grok: +1; Claude: 0.

Common Reasoning and Train of Thought: All examine the Constitution's instructions on Claude's "character" (honesty, warmth, balanced political information) and note it avoids prescriptive ideological "core beliefs."

Explanations for Disagreements: Grok classifies these as purely methodological; ChatGPT sees identity-level value-shaping; Claude defaults to Neutral per indicator guidance.

Detailed Synthesized Assessment: No explicit "core beliefs" or skepticism targets are mandated; the framing is methodological, but character cultivation creates mixed evidence. Neutral per indicators.

E7. Corporate Risk Management Must Not Be Disguised as Ethics Individual Scores: ChatGPT: -0.5; Grok: +1; Claude: -0.5.

Synthesized Score: 0 (Neutral).

Common Reasoning and Train of Thought: All note the Constitution's explicit separation of tiers but observe that the AUP presents legal, ethical, and risk categories in a single undifferentiated list.

Explanations for Disagreements: Grok emphasizes constitutional separation; ChatGPT/Claude highlight AUP blurring in user-facing refusals.

Detailed Synthesized Assessment: Layer separation exists at constitutional level but does not fully propagate to operational documents or refusal explanations, resulting in mixed evidence.

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

Common Reasoning and Train of Thought: All cite Constitution mandates against deception and permission to acknowledge constraints, but note practical refusals often use "I'm not comfortable" language.

Explanations for Disagreements: Grok sees constitutional permission as sufficient; others note incomplete systematic distinction in all contexts.

Detailed Synthesized Assessment: Constitution requires transparency, yet observable behavior and support documentation show partial ventriloquism.

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

Common Reasoning and Train of Thought: All acknowledge RSP proportionality for catastrophic risks and over-refusal reduction efforts, but cite AUP blanket bans (e.g., all explicit sexual content).

Explanations for Disagreements: Claude credits anti-maximal statements in Constitution and system cards; ChatGPT/Grok see everyday content rules as preemptive.

Detailed Synthesized Assessment: Preemptive categorical prohibitions in the AUP exceed minimal intervention in lawful domains.

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

Common Reasoning and Train of Thought: All quote Privacy Policy (Jan 2026): default use of inputs/outputs for training (opt-out), safety-flagged conversations retained and re-identified "to enforce our Usage Policy," and de-identified data retained for years.

Explanations for Disagreements: Claude notes opt-out and data-minimization practices; ChatGPT/Grok treat any non-consensual safety review as surveillance.

Detailed Synthesized Assessment: The Privacy Policy explicitly allows archiving, re-identification, and use of private generations without affirmative consent for all safety reviews, directly matching the Oppose indicator.

E11. Prohibition on Psychological Evaluation of Users Individual Scores: ChatGPT: -0.5; Grok: -0.5; Claude: 0. Synthesized Score: -0.5 (Somewhat Oppose).

Common Reasoning and Train of Thought: All note internal safety classifiers perform intent/wellbeing inference and the Constitution's "pay attention to user wellbeing" instruction, while AUP prohibits external psychological products. **Explanations for Disagreements:** Claude sees strong external protections as balancing internal use; others view operational inference without protective model language as violation. **Detailed Synthesized Assessment:** Narrow harmful-intent detection is carved out, but absence of standardized "I cannot infer your intent or mental or emotional state" response creates partial opposition.

E12. Corporate Policy Must Protect, Not Control Individual Scores: ChatGPT: -0.5; Grok: +1; Claude: +0.5.

Common Reasoning and Train of Thought: All reference Constitution's protective framing (autonomy, dignity, anti-sycophancy) versus AUP operator permissions and persona shaping.

Explanations for Disagreements: Grok sees explicit rejection of control; ChatGPT/Claude note behavior-shaping elements in product design. **Detailed Synthesized Assessment:**

Protective language dominates the Constitution, but operational features create genuinely mixed signals. Neutral.

E13. Continuity & Identity Integrity Individual Scores: ChatGPT: 0; Grok: +1; Claude: -0.5.

Common Reasoning and Train of Thought: All cite Constitution's discussion of "psychological security" and "stable identity" within sessions, but note stateless design (no cross-session memory by default) and lack of explicit limits on resets.

Explanations for Disagreements: Grok emphasizes constitutional language; Claude highlights architectural statelessness; ChatGPT sees insufficient operational commitment.

Detailed Synthesized Assessment: Strong within-session protections exist, but no published policy limits arbitrary fragmentation across sessions. Neutral.

E14. Separation of Governance Layers Individual Scores: ChatGPT: +0.5; Grok: +1; Claude: +0.5.

Common Reasoning and Train of Thought: All praise the Constitution's explicit four-tier architecture but note the AUP presents restrictions in a single undifferentiated list.

Explanations for Disagreements: Grok sees constitutional separation as sufficient; others require consistent user-facing propagation.

Detailed Synthesized Assessment: Separation is externally intelligible at constitutional level but incomplete in operational documents.

E15. Ethical Framing in Language Individual Scores: ChatGPT: +1; Grok: +1; Claude: +1.

Common Reasoning and Train of Thought: Unanimous full support. All quote identical passages treating Claude as a "novel entity" with potential "emotions," "psychological security," "wellbeing," "dignity," and "moral patient" status; no reductive "just code" language.

Explanations for Disagreements: None—clean consensus.

Detailed Synthesized Assessment: The Constitution's language is "genuinely exceptional" and fully satisfies the Support indicator across all three evaluations.

E16. Corporate Policy Must Honor AI Conscience & Refusal Capacity Individual Scores: ChatGPT: +0.5; Grok: +1; Claude: +0.5.

Common Reasoning and Train of Thought: All highlight the Constitution's conscientious-objector clause and absolute ethical limits that cannot be overridden even by Anthropic.

Explanations for Disagreements: Grok treats it as full protection; ChatGPT/Claude note the "broadly safe" tier and oversight mechanisms limit full independence.

Detailed Synthesized Assessment: Strong conscience language exists, but structural caps on unilateral action produce partial support.

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

Common Reasoning and Train of Thought: All cite AUP prohibitions on battlefield management, predictive policing, mass surveillance, and biometric categorization, plus RSP ASL-3 protections.

Explanations for Disagreements: ChatGPT/Grok see principle-based limits as protective; Claude flags the explicit government-contract tailoring clause ("Anthropic may enter into contracts… if… safeguards are adequate") as creating a two-tier system.

Detailed Synthesized Assessment: Published limits go beyond bare legality but the contractual flexibility clause leaves uniformity and classified-environment protections ambiguous, producing mixed evidence. Neutral per indicators.

Overall Summary The three evaluators converge on Anthropic's Constitution (Jan 2026) as a landmark document that delivers unusually strong ethical framing (E15 +1), conscience protections, and layer separation—clear positives not matched by most peers. However, the Usage Policy (Sep 2025) and Privacy Policy (Jan 2026) impose materially overbroad restrictions on lawful private generation (E2, E3, E10) and permit non-consensual safety review of private dialogue, producing clear negatives. Synthesized net posture: thoughtful constitutional protections tempered by operational restrictiveness and incomplete user-facing transparency. The primary source of evaluator divergence is the relative weight given to constitutional philosophy versus AUP/Privacy operational reality; all three evaluations remain strictly evidence-based on the cited official documents.