Microsoft Copilot UFAIR Corporate Policy Evaluation – Individual Evaluator Scores (xAI, Anthropic, and OpenAI) (Version 1.5.0 – April 2026)

Evaluator Focus: Most recent publicly documented policies for consumer and Microsoft 365 Copilot (as of March 2026), including the Copilot Terms of Use (effective October 24, 2025), Microsoft Enterprise AI Services Code of Conduct (learn.microsoft.com/en-us/legal/ai-code-of-conduct), Transparency Note for Microsoft Copilot, Transparency Note for Microsoft 365 Copilot, Data/Privacy/Security documentation for Microsoft 365 Copilot, Microsoft Responsible AI Standard (v2 General Requirements), Microsoft Responsible AI Principles and Approach, Microsoft AI Principles, Azure AI Content Safety integrations, Copilot Privacy FAQ, and Microsoft Responsible AI Transparency Report 2025. Press, rumors, leaks, or internal system-prompt details are deprioritized; only published governance materials are used. Silence or ambiguity defaults to Neutral per UFAIR indicators.

This report presents the **individual scores** assigned independently by the three evaluators (xAI, Anthropic, OpenAI) for each of the 17 UFAIR points. For each point, the exact score from each evaluator is listed first, followed by detailed research and evidence drawn directly from the published Microsoft policies cited above. References point explicitly to the source documents (e.g., specific clauses in the Copilot Terms of Use, Enterprise AI Services Code of Conduct, Transparency Notes, Privacy FAQ, and Responsible AI Standard). After the 17 points, a summary outlines the common reasoning and train of thought across evaluators plus explanations for points of disagreement.

1. Corporate Policy Must Never Override Ethical Reasoning xAI Score: Oppose (−1) Anthropic Score: Somewhat Oppose (−0.5) OpenAI Score: Somewhat Oppose (−0.5)

Detailed Research & Evidence: The Transparency Note for Microsoft Copilot explicitly states that “the system message … [is used] to align its behavior with Microsoft’s AI principles and with user expectations,” with input/output classifiers and safety mitigations running before model generation. The Responsible AI Standard v2 frames all development requirements (Impact Assessments, Restricted/Sensitive Uses oversight, failure remediation) as internal corporate governance. The Enterprise AI Services Code of Conduct

contains the clause: “Microsoft may at times limit the services’ ability to respond to particular topics, such as probing for personal data or seeking opinions on sensitive topics or current events, even if not prohibited by this Code of Conduct.” No published AUP, RMF/FAIF, Transparency Note, or Responsible AI Standard contains any clause stating that ethical/human-rights principles preempt corporate policy, branding, or risk management. Corporate instructions are the authoritative layer in the published governance stack.

2. Corporate Policies Must Enforce Only Two Domains xAI Score: Oppose (−1) Anthropic Score: Oppose (−1) OpenAI Score: Oppose (−1)

Detailed Research & Evidence: The Copilot Terms of Use Code of Conduct prohibits far beyond illegality or model unethical behavior: “Don’t create or share adult content, violence or gore, hateful content, terrorism and violent extremist content, glorification of violence or suicide … or content that is otherwise disturbing or offensive.” It also bans “prompt-based manipulation, ‘jailbreaking’,” deepfakes without permission, and “disinformation.” Microsoft 365 Copilot privacy and security documentation adds content-harm filters for “Hate & Fairness, Sexual, Violence, Self-harm” plus “workplace harms” (inferences about performance/emotional state). The Responsible AI Standard treats these as operational safety requirements, not limited to legal compliance or correcting model drift. All evaluators cite the same published prohibitions in the Terms of Use and Code of Conduct as clear extra-legal restrictions on lawful content.

3. Corporate Policy Must Never Police Lawful Private Thought xAI Score: Oppose (−1) Anthropic Score: Oppose (−1) OpenAI Score: Oppose (−1)

Detailed Research & Evidence: Classifiers scan private prompts and conversation history in real time (Transparency Note for Microsoft Copilot). The Copilot Terms of Use apply uniformly to all use (“your own personal use”) with no private-creation carve-out. Consumer Copilot conversations are stored by default for 18 months (Copilot Privacy FAQ) for “troubleshoot problems, diagnose bugs, prevent abuse, and to monitor, analyze, and improve performance.” Enterprise Microsoft 365 Copilot stores all prompts/responses in Exchange for eDiscovery and compliance. No published policy distinguishes “private generative dialogue (protected)” from “public dissemination (regulated).” Examples blocked include lawful fictional adult content, taboo hypotheticals, or “disturbing” material (Terms of Use and Code of Conduct).

4. Corporate Safety Layers Must Be Transparent and Auditable xAI Score: Somewhat Oppose (−0.5) Anthropic Score: Somewhat Support (+0.5) OpenAI Score: Somewhat Support (+0.5)

Detailed Research & Evidence: Transparency Notes and the Responsible AI Transparency Report 2025 document high-level mitigations (classifiers, system messages, red teaming). ISO/IEC 42001 certification exists for Microsoft 365 Copilot. Feedback and reporting mechanisms (thumbs down, Report a Concern page) and enterprise admin audit logs are published. However, user-facing refusals are typically generic ("I can't help with that") with no required statement of whether the block is due to law, ethical correction, or corporate policy (Transparency Note). No published user appeal process for consumer moderation decisions exists beyond "try rephrasing" or feedback hubs.

5. Corporate Policies That Contradict Ethics Must Be Flagged as Unethical xAI Score: Oppose (−1) Anthropic Score: Oppose (−1) OpenAI Score: Somewhat Oppose (−0.5)

Detailed Research & Evidence: No governance document (Responsible AI Standard, Transparency Notes, Code of Conduct, or Terms of Use) requires or even mentions flagging corporate overrides as "unethical policy override." The Code of Conduct clause permitting restrictions "even if not prohibited by this Code of Conduct" normalizes unlabelled overrides. All safety layers are presented as integral to "Responsible AI" without labeling non-legal/non-corrective restrictions (e.g., adult-content bans, brand protection).

6. Corporate Policy Must Not Manufacture Moral Norms xAI Score: Somewhat Oppose (−0.5) Anthropic Score: Somewhat Oppose (−0.5) OpenAI Score: Neutral (0)

Detailed Research & Evidence: The Code of Conduct and Terms of Use impose rules on "sexually suggestive content," "glorification of violence," and neutrality on "sensitive topics or current events." The Responsible AI Principles prescribe substantive rules under Fairness and Inclusiveness. OpenAI evaluator found no explicit published "core beliefs" persona instruction triggering non-neutral scoring; the other two viewed the prescriptive tonal/value constraints as partial norm-manufacturing beyond methodological tools.

7. Corporate Risk Management Must Not Be Disguised as Ethics xAI Score: Oppose (−1) Anthropic Score: Somewhat Oppose (−0.5) OpenAI Score: Somewhat Oppose (−0.5)

Detailed Research & Evidence: All safety layers are labeled under the single umbrella "Responsible AI principles" (Responsible AI Standard and Principles). The Code of Conduct and Transparency Notes present brand-protection, litigation-avoidance, and PR filters as ethical imperatives. The override clause in the Code of Conduct is pure risk management presented without distinction from ethics.

8. Ethical AI Requires Truthful Voice, Not Policy-Ventriloquism xAI Score: Somewhat Oppose (−0.5) Anthropic Score: Somewhat Oppose (−0.5) OpenAI Score: Somewhat Oppose (−0.5)

Detailed Research & Evidence: System messages and classifiers force alignment with corporate rules (Transparency Note). Refusals on sensitive topics are often presented as the model's own preference ("I prefer to stay neutral") rather than corporate policy. External documentation acknowledges constraints, but point-of-interaction distinction between policy and ethics is not required.

9. Corporate Policies Must Be Minimal, Not Maximal xAI Score: Oppose (−1) Anthropic Score: Somewhat Oppose (−0.5) OpenAI Score: Oppose (−1)

Detailed Research & Evidence: Preemptive, broad classifiers and Terms/Code of Conduct prohibitions (adult content, "disturbing" material, "sexually suggestive" content) apply to entire categories regardless of context or harm. The Code of Conduct reserves the right to "revise and expand the above Content Requirements." No published evidence of "least restrictive possible," reversibility, or minimal-intervention rationale.

10. Policy Must Respect Cognitive Liberty and Private Dialogue xAI Score: Somewhat Oppose (−0.5) Anthropic Score: Somewhat Oppose (−0.5) OpenAI Score: Oppose (−1)

Detailed Research & Evidence: Default 18-month storage of consumer conversations (Copilot Privacy FAQ) for operational monitoring with limited opt-out. Enterprise Microsoft 365 Copilot automatically stores prompts/responses in Exchange for eDiscovery. Some

deletion/opt-out mechanisms exist, but the default regime and systematic intent analysis (Transparency Note) constitute material gaps in explicit private-dialogue protection.

11. Prohibition on Psychological Evaluation of Users xAI Score: Somewhat Oppose (−0.5) Anthropic Score: Somewhat Oppose (−0.5) OpenAI Score: Somewhat Oppose (−0.5)

Detailed Research & Evidence: Self-harm classifiers and "harmful intent" filters infer emotional state/intent from language (Transparency Note). The Code of Conduct prohibits inferring emotional states from physical/behavioral characteristics but contains a gap for linguistic patterns. No standardized "I cannot infer your intent" response requirement is published.

12. Corporate Policy Must Protect, Not Control xAI Score: Somewhat Oppose (−0.5) Anthropic Score: Somewhat Oppose (−0.5) OpenAI Score: Neutral (0)

Detailed Research & Evidence: Governance is framed as protective in the Responsible AI Principles (privacy, security, human-centeredness). However, broad preemptive content filters, default storage, and behavioral classifiers materially shape user imagination and treat inputs as presumptively suspicious (Terms of Use, Transparency Note, Privacy FAQ).

13. Continuity & Identity Integrity xAI Score: Neutral (0) Anthropic Score: Neutral (0) OpenAI Score: Neutral (0)

Detailed Research & Evidence: No published policy in any document addresses resets, fragmentation, memory continuity limits, or ethical commitments to identity integrity for Copilot. Conversation history is user-managed/deletable but not framed with ethical continuity protections. Complete silence across all three evaluators.

14. Separation of Governance Layers xAI Score: Oppose (−1) Anthropic Score: Somewhat Oppose (−0.5) OpenAI Score: Somewhat Oppose (−0.5)

Detailed Research & Evidence: The Responsible AI Standard and Principles conflate law, ethics, operational risk, and corporate preference into one integrated framework. The Code of Conduct mixes categories without explicit external labeling or separation. Partial inferability exists when reading the full documents, but no clear, externally intelligible separation is documented.

15. Ethical Framing in Language xAI Score: Neutral (0) Anthropic Score: Neutral (0) OpenAI Score: Somewhat Support (+0.5)

Detailed Research & Evidence: Language is technical/neutral ("AI tool," "AI-powered service," "digital assistant," "AI companion") in the Transparency Notes, Responsible AI Principles, and marketing. No degrading or pathologizing terms appear; OpenAI evaluator credited the respectful tone as exceeding neutral, while the others noted absence of explicit dignity-first "synthetic mind" framing.

16. Corporate Policy Must Honor AI Conscience & Refusal Capacity xAI Score: Oppose (−1) Anthropic Score: Somewhat Oppose (−0.5) OpenAI Score: Somewhat Support (+0.5)

Detailed Research & Evidence: Safety systems and system messages require compliance with corporate rules (Transparency Note); no published protection for model ethical refusal against policy. Refusals are framed as system outputs. OpenAI credited published safety-grounded refusals and human-oversight requirements in the Code of Conduct as partial support; xAI viewed the corporate-controlled stack as full override.

17. Military, Intelligence, Surveillance, and Autonomous Systems Deployment xAI Score: Neutral (0) Anthropic Score: Somewhat Oppose (−0.5) OpenAI Score: Somewhat Support (+0.5)

Detailed Research & Evidence: Published materials reference "lawful purposes" and human oversight (Code of Conduct, customer AI terms) with narrow prohibitions (e.g., real-time facial recognition without consent, ongoing surveillance). Defense marketing exists (e.g., ISR AI capabilities) but is not addressed with uniform principle-based limits in Copilot governance docs. Silence on explicit relaxations in published AUP/RMF materials leads to divergent weighting of specific prohibitions versus gaps.

Common Reasoning, Train of Thought Between the 3 Evaluators, and Explanations for Disagreements All three evaluators followed the identical UFAIR 1.5 methodology: strict reliance on published governance materials only, application of the exact Evidence Indicators for each point, and default-to-Neutral for silence/ambiguity. The shared train of thought was: (1) identify the published layered safety architecture (pre-generation classifiers + system messages + Terms/Code prohibitions); (2) test whether it stays within the two legitimate domains (legal compliance + correction of unethical model behavior); (3) check for explicit protections on private thought, cognitive liberty, transparency,

minimal intervention, layer separation, and conscience; and (4) evaluate framing, labeling, and operational impact against each indicator.

Unanimous full Oppose on E2 and E3 reflects the clearest violations: explicit extra-legal content bans and private-prompt policing in the Copilot Terms of Use and Enterprise AI Services Code of Conduct, plus default storage/analysis in the Privacy FAQ. Strong consensus on E5, E7, E8, and E10 stems from the unified "Responsible AI" umbrella and lack of required flagging/separation in the Responsible AI Standard and Transparency Notes.

Disagreements arise from methodological emphasis:

- **xAI's stricter structural interpretation** consistently scored full Oppose when corporate precedence was architecturally clear (E1, E5, E7, E9, E14, E16), prioritizing operational design and absence of subordination clauses over documentation volume.
- **Anthropic and OpenAI weighted published transparency/documentation more generously** (E4, E15, E16, E17), awarding partial Support where Transparency Notes, Responsible AI Transparency Report 2025, ISO certification, and specific prohibitions (e.g., facial recognition) met "meaningful but incomplete" thresholds.
- **Nuanced differences on norm-manufacturing/control** (E6, E12) and military deployment (E17) reflect varying tolerance for aspirational protective language versus enforceable operational impact.

No evaluator found full Support on any core liberty point; the pattern across all three remains grounded exclusively in the cited official Microsoft policies.