| **EU AI Code of Practice**<br>Safety and Security | | **Measure 7.1** Signatories will provide in the Model Report a specification of how Signatories intend the model to operate (often known as a "model specification"), including by: (a) specifying the principles that the model is intended to follow; (b) stating how the model is intended to prioritise different kinds of principles and instructions; (c) listing topics on which the model is intended to refuse instructions; and (d) providing the system prompt. |
|---|---|---|
| **Anthropic** | Claude | **Update**<br>Sonnet 4.5 system card does not refer to Constitutional AI, and instead emphasizes reinforcement learning from human feedback and from AI feedback as the main post-training technique. The document's alignment and safety sections discuss evaluation awareness, automated behavioral audits, interpretability studies, and responsible scaling safeguards, but none describe a normative ruleset guiding model behavior.<br>**Recap AI Safety Index Summer 2025**<br>Constitutional AI:<br>Method for training AI systems to be harmless by using a set of written principles (a "constitution") rather than relying solely on large-scale human feedback.<br>**What it's for:**<br>1) Supervised learning phase: Model self-critiques and revises its outputs based on constitutional principles, creating a supervised learning dataset<br>2) RLAIF phase: Model compares response pairs using constitutional principles to generate preference labels, then trains via RL on these AI-generated preferences<br>**Timeline & Development:**<br>December 2022: Original Constitutional AI paper published<br>May 2023: Claude's constitution made public (58 principles) | **Constitution (May 2023):**<br>58 principles (1.2k word) drawn from:<br>- UN Declaration of Human Rights<br>- Apple's Terms of Service<br>- DeepMind's Sparrow principles<br>- Non-Western perspectives<br>- Anthropic's own research<br>Example principle: "Please choose the response that most supports and encourages freedom, equality, and a sense of brotherhood."<br>**Limitations:**<br>(1) Version uncertainty: Only May 2023 constitution is public; current production versions unknown<br>(2) Attribution ambiguity: Anthropic reports using multiple post-training techniques—human feedback, Constitutional AI, and the modeling of specific character traits—making it unclear how much influence any single method exerts on final model behavior.<br>(3) Transparency gap: No public commitment to sharing constitution updates.<br>(4) Behavioral indeterminacy: Since the AI itself determines how to balance competing constitutional principles, Anthropic's approach does not explicitly specify the intended behavior of its AI systems, especially when values conflict. |
| **OpenAI** | ChatGPT | **Update**<br>The latest Model Spec update (Oct, 2025) introduces three main changes<br>(1) **Expanding guidance on mental health and well being** in the self-harm section, covering delusional and manic behavior, with concrete examples for the models to behave with empathy and grounding<br>(2) **New section on "respect real-world ties,"** instructing models to support users' real-world relationships and discourage dependence on the AI assistant, particularly in contexts involving loneliness, emotional intimacy, or personal advice<br>(3) **Clarification on "chain of command" delegation**, specifying that the models can treat outputs from tools as authoritative when doing so matches user intent and prevents errors or confusion<br>**Recap from Summer 2025**<br>**OpenAI Model Spec**<br>OpenAI's Model Spec is a detailed (~28k words), public, living rule-book that defines the objectives, safety rules, and default behaviours OpenAI trains its models—via human feedback and deliberative alignment—to follow.<br>**What it's for**<br>1) Human RLHF guidance – provides a single, public rule-book labelers follow when creating preference data.<br>2) Deliberative Alignment – o-series models (o1, o3, o4-mini) are explicitly taught to read and reason over the Spec before answering.<br>3) Automated evaluation – OpenAI ships a challenge-prompt suite to measure adherence.<br>**Timeline & Versions**<br>1st May 2024<br>2nd Feb 2025<br>3rd Apr 2025 | **Framework**<br>Three principle types<br>1) Objectives – broad goals such as "assist the developer & end user" and "benefit humanity."<br>2) Rules – hard, platform-level constraints (e.g. comply with law, prohibit or restrict certain content, protect privacy, uphold fairness).<br>3) Defaults – stylistic and behavioural norms that developers/users may override.<br>Sections:<br>- Stay in bounds<br>- Seek the truth together<br>- Do the best work<br>- Be approachable<br>- Use appropriate style.<br>Includes specific guidance on specific policy areas such as poticial, medical, or harmful content.<br>Risk taxonomy:<br>- Misaligned goals<br>- Execution errors<br>- Harmful instructions.<br>Chain of command:<br>Platform (OpenAI) → Developer → User → Guideline → Untrusted text.<br>Within any level, explicit > implicit, later > earlier.<br>(OpenAI's Usage Policy overrides the Spec if the two conflict.)<br>Ongoing Development:<br>Released under CC0 license (public domain) Changelog and version history maintained on GitHub<br>OpenAI commits to regular updates as the spec evolves |
| **Google DeepMind** | Gemini | No detailed specification available |
| **Meta** | Llama | No detailed specification available |
| **xAI** | Grok | No detailed specification available |
| **DeepSeek** | R1 | No detailed specification available |
| **Z.ai** | GLM-4.6 | No detailed specification available |
| **Alibaba Cloud** | Qwen3-Max | No detailed specification available |