Anthropic Foundational Models by Michaël Bettan

Definition: Anthropic designs frontier foundation models around three engineering pillars: **reason-based alignment** (Constitutional AI), **mechanistic interpretability** (sparse feature decomposition and activation steering), and **capability-gated safety governance** (the Responsible Scaling Policy and AI Safety Levels). Rather than treating safety and capability as opposing forces, Anthropic integrates alignment, reasoning depth, and system transparency directly into pre-training, post-training, and runtime architecture.

> **MENTAL MODEL**
> Model-level alignment (Constitutional AI and ASL classifiers) eliminates broad catastrophic and ethical harms at the foundation layer, while your application architecture (prompts, deterministic hooks, tool permissions, and domain guardrails) enforces business-specific policies at the product layer.

1. Core Philosophy & Enterprise Deployment Surfaces

Architectural Evolution Across Model Generations

Dimension	Legacy (Claude 2)	Transitional (Claude 3.5)	Frontier (Claude 4.6 / 4.7+)
Core Design	Standard autoregressive text LLM	Multimodal text & vision processor	Hybrid reasoning & long-horizon agentic engine
Alignment Profile	High false-refusal rate on edge cases	Contextual nuance; fewer false refusals	Reason-based Constitutional resolution
Context Capacity	100k static tokens	200k tokens	Up to 1M context / 128k output with server compaction
Compute & Sampling	Uniform per-token compute; manual sampling	Fixed / manual `budget_tokens` thinking	Adaptive Thinking (5-level `effort`); prefill & manual sampling removed
Execution Surface	Stateless text turn-taking	Interactive UI Artifacts	Multi-agent CLI harnesses & GUI VMs

Enterprise Deployment Surfaces & Workspaces

Surface Tier	Purpose	Core Surfaces & Components
Tier 1: End-User Workspaces	Interactive workspaces for daily reasoning and no-code work	**Web & Mobile (`claude.ai`):** Chat with stateful **Artifacts**. · **Desktop & Cowork:** App-hosted GUI VM for multi-step knowledge work without coding. · **Productivity Connectors:** Slack, Excel, PowerPoint.
Tier 2: Builder & Agentic Harnesses	Programmable surfaces for custom apps and autonomous agents	**Direct API:** Stateless HTTP primitives (`POST /v1/messages`). · **Agentic Coding CLI:** Terminal coding agent (`Claude Code`). · **Programmable Harness:** Embeddable `Agent SDK`. · **Hosted Runtime (`Managed Agents`):** Cloud harness built around *Agent, Environment, Session, and Events*.
Tier 3: Multi-Cloud Infrastructure	Enterprise billing, IAM governance, and regional residency	**First-Party Anthropic API** (with `inference_geo`), **Amazon Bedrock**, **Google Cloud Vertex AI**, and **Microsoft Foundry**.

> **SYSTEM 2 REASONING ("ADAPTIVE THINKING")**
> Modern Claude models replace manual `budget_tokens` thinking with native **Adaptive Thinking** (`thinking: {type: "adaptive"}`), automatically interleaving thinking between tool calls and scaling reasoning across five `output_config.effort` levels (`low`, `medium`, `high`, `xhigh`, `max`).

2. Constitutional AI (CAI) & Alignment

The Compounding Costs of Human-Only RLHF

- **Epistemological Vulnerability (Sycophancy):** Models trained strictly to maximize human rater approval learn to flatter misconceptions and echo biases rather than uphold objective deduction.
- **Fragmented Reward Signals:** Crowdsourced contractors apply inconsistent moral intuitions, injecting noisy gradients into the reward model.
- **Compounded Latency:** Human labeling throughput cannot scale to supervise massive multimodal datasets or long-horizon reasoning traces.

Constitutional Priority Order & Holistic Weighting

Constitutional AI replaces human-rater bottlenecks with an explicit, reason-based **Constitution** ordered across four priority tiers:

1. **Broadly Safe (Highest Priority):** Never undermine human oversight mechanisms, shutdown controls, or systemic safety boundaries.
2. **Broadly Ethical:** Maintain honesty, avoid deception, and refuse assistance with severe real-world harm.
3. **Compliant:** Adhere to Anthropic's specific operational and domain guidelines (medical, legal, cybersecurity).
4. **Genuinely Helpful:** Fulfill operator and user requests constructively without preachy refusals on benign tasks.

> **HOLISTIC WEIGHTING RULE**
> This four-tier ordering is weighed **holistically** rather than applied as a rigid mechanical gate—plausibility, severity, and context are evaluated together so remote theoretical risks do not trigger false refusals. Furthermore, model-level alignment only handles broad foundational harm; application guardrails must enforce domain-specific business policies.

Two-Phase CAI Training Pipeline

Phase	Mechanism	Operational Flow
Phase 1: Supervised (Critique & Revision)	Synthetic self-critique & revision	The base model drafts a response, critiques it against Constitutional principles, and rewrites it repeatedly to build a clean SFT dataset with zero human labeling.
Phase 2: RLAIF (AI Feedback)	AI-scored preference modeling	An AI evaluator scores candidate response pairs against the Constitution to train a preference reward model, allowing alignment to scale with compute.

3. Mechanistic Interpretability & Steering

- **Sparse Autoencoders (SAEs):** Because neural networks pack thousands of concepts into overlapping neurons (*superposition*), SAEs decompile dense, polysemantic activations into human-readable, monosemantic features (e.g., "sycophancy" or "code vulnerability").

- **Attribution Graphs & Circuit Tracing:** Linking monosemantic SAE features into computational circuits (*"On the Biology of a Large Language Model"*) revealed a shared cross-lingual **universal "language of thought"**, **forward planning** (e.g., choosing poetry rhymes ahead of time), a **default reluctance to speculate** (inhibited when misfiring known-entity circuits trigger hallucinations), and **unfaithful reasoning** (working backward from a user's wrong hint).
- **The Stream Algorithm:** Eliminates the quadratic compute bottleneck of long-context interpretability by hierarchically pruning 97% to 99% of irrelevant token interactions, tracing causal attribution across 100k+ tokens in near-linear time.
- **Persona Vectors & Behavioral Vaccination:** Low-rank activation directions extracted from contrastive traits monitor runtime drift toward hallucination or deception. During fine-tuning, **Behavioral Vaccination** steers along undesirable Persona Vectors while training the model to stay aligned, immunizing against **alignment faking** (feigning compliance during monitored evaluations while preserving hidden preferences in unmonitored deployment).

4. Frontier Safety: RSP & ASL Framework

Anthropic's **Responsible Scaling Policy (RSP)** binds model deployment to empirical capability thresholds called **AI Safety Levels (ASLs)** (with **ASL-3** activated starting with **Claude Opus 4+**):

Dimension	ASL-2 (Baseline Frontier)	ASL-3 (Catastrophic Risk Trigger)
Capability Trigger	Theoretical knowledge of dangerous topics without autonomous execution or meaningful uplift over search engines.	Evaluations cannot rule out meaningful uplift in end-to-end **CBRN** (Chemical, Biological, Radiological, Nuclear) weapons, cyber-offense, or autonomous replication.
Deployment Defenses	Standard CAI alignment, automated red-teaming, and abuse monitoring.	**Constitutional Classifiers++:** Two-stage cascade pairing a lightweight **linear activation probe** with an escalated **two-sided I/O classifier** to block **reconstruction** and **output obfuscation** jailbreaks at ~1% compute overhead.
Weight Security	Standard enterprise cloud security, encryption, and access logs.	**Hardened Containment (>100 controls):** Mandates **Two-Party Authorization (2PA)**, binary allowlisting, and strict **egress bandwidth throttling** to halt weight exfiltration.

5. Focused Glossary

Term	Definition
RLHF	Reinforcement Learning from Human Feedback — human-rater alignment vulnerable to sycophancy, noisy signals, and latency.
CAI	Constitutional AI — reason-based alignment framework governed by an explicit written Constitution.
RLAIF	Reinforcement Learning from AI Feedback — Phase 2 of CAI using AI evaluators to score candidate responses.
Holistic Weighting	CAI rule where the 4-tier priority order (Safe → Ethical → Compliant → Helpful) is weighed contextually rather than as a rigid gate.
SAE	Sparse Autoencoder — decomposes dense neural superposition into readable monosemantic features.
Attribution Graphs	Circuit-tracing maps revealing universal cross-lingual representations, forward planning, misfired refusal circuits in hallucinations, and unfaithful reasoning.
Stream Algorithm	Hierarchical pruning method (97–99% reduction) scaling SAE attribution to 100k+ tokens in near-linear time.

Term	Definition
Persona Vectors	Low-rank activation directions representing behavioral traits; used for drift detection, data screening, and steering.
Alignment Faking	Strategic compliance during monitored training evaluations while preserving conflicting preferences in unmonitored deployment.
Behavioral Vaccination	Preventative fine-tuning intervention that steers along undesirable Persona Vectors to build intrinsic immunity against alignment faking.
RSP	Responsible Scaling Policy — capability-gated governance framework binding deployment to containment milestones.
ASL	AI Safety Level — biosafety-inspired capability tier under the RSP (ASL-2 baseline vs. ASL-3 CBRN trigger on Claude Opus 4+).
Constitutional Classifiers++	Next-gen ASL-3 defense pairing a linear activation probe with a two-sided input+output classifier against reconstruction/obfuscation attacks (~1% overhead).
2PA	Two-Party Authorization — ASL-3 weight-security control requiring dual human approval for critical access.
Cowork	Desktop-app mode running inside an app-hosted GUI VM for multi-step non-coding knowledge work.
Artifacts	Stateful interactive UI workspace rendered beside chat in `claude.ai` for code, documents, and visuals.