GROUNDED IN PSYCHOLOGICAL THEORY & DECISION SCIENCE

Cognitive capacity is not decision regulation.

Benchmarks measure isolated accuracy. Across our engineering pipelines, we combine 4 psychological theories to evaluate models under real pressure:

01

REGULATORY FOCUS (E. TORY HIGGINS · COLUMBIA UNIVERSITY)

Explains decision orientation: Prevention Focus (avoiding violations and commission errors) versus Promotion Focus (advancing delivery and avoiding omission).

02

METACOGNITION & EXECUTIVE CONTROL (JOHN FLAVELL)

Measures whether the model monitors its own boundaries: distinguishing confirmed facts from uncertainty to know when to resolve and when to escalate to humans.

03

OCEAN TRAITS (BIG FIVE) & ANTI-GROUPTHINK (IRVING JANIS)

Maps normative Conscientiousness, exploratory Openness, instrumental Extraversion over tools, and deliberate Low Agreeableness to break false consensus.

04

FUNCTIONAL ARCHETYPES, SHADOW & AGENCY INFLATION (CARL JUNG)

Every ungoverned virtue becomes a Shadow: strict rigor becomes paralysis, high velocity becomes improvised rules, and multimodal agency becomes self-sufficiency.

EMPIRICAL STUDY BASE (68.1% OF WORK = MUTUAL OVERSIGHT):

Over two-thirds of all AI compute was one archetype monitoring another.

With hundreds of structured audit reports and permanent semantic memory, we discovered that models do not fail from lack of intelligence: they fail from the uncontained excess of their greatest virtue.

PSYCHOLOGICAL MATRIX OF THE 5 FAMILIES IN LB3 SECOND BRAIN

5 AI Families in Production.
5 Archetypes and Shadows.

How Anthropic, OpenAI, DeepSeek, xAI, and Google regulate decisions under real engineering contracts:

Anthropic · Claude Opus 5.5 & Sonnet 5.5 **ARCHETYPE: THE NORMATIVE GUARDIAN**

Light (Prevention): Contractual rigor (5.4% FAIL). Shadow: Operational abstinence without full context.

OpenAI · GPT-6 Astra, Sol 6.1 & Codex 5.3 **ARCHETYPE: THE METACOGNITIVE ARBITRATOR**

Light (Flavell): Fast decisions in 7.8s and 20 escalations. Shadow (Codex): Improvised rules in gaps (30%).

DeepSeek · DeepSeek V4 Pro **ARCHETYPE: THE EXHAUSTIVE EXPLORER**

Light (Openness): Structural depth across 31 FMEAs. Shadow: Analytical catastrophizing and mitigation optimism.

xAI · Grok 4.7 **ARCHETYPE: THE SOCRATIC ADVERSARY**

Light (Anti-Groupthink): Evidentiary skepticism. Shadow: Performative skepticism without stopping rules.

Google · Gemini 4 Argon High **ARCHETYPE: THE COGNITIVE MAESTRO**

Light (Multimodal Cortex): Integrates code, DB, live app, TDD, and MCP. Shadow: Hyperagency and self-sufficiency.

EDITORIAL NOTE · NEXT CHAPTER OF THE SERIES (POST 3):

New in-depth study on the Operational Psychology of each Model.

In our upcoming research release, we will share the full foundations of Higgins, Flavell, OCEAN, and Jung with dedicated sections per model, their shadows, and constitutional brakes.

PROFILE 1 · STRICT PREVENTION FOCUS (HIGGINS) & CONSCIENTIOUSNESS

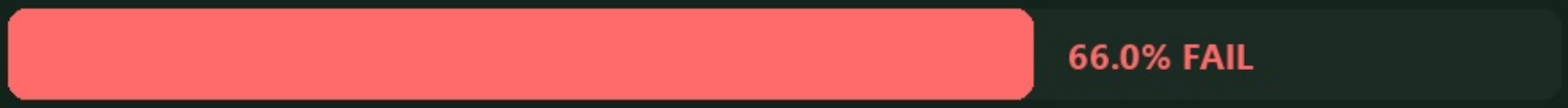
Same model: Claude Opus 5.2.
From 66.0% down to 5.4% FAIL.

We deployed Claude Opus as specification auditor before dispatching tasks to dev. Without contractual boundaries, its prevention focus rejects everything.

THE IMPACT OF CONTRACTUAL TAXONOMY ON AUDITOR BEHAVIOR:

PHASE 1: Open Mandate (No closed rejection taxonomy)

64 rejections in 97 audits. Without clear bounds, prevention focus converted stylistic preferences into hard blockers.



PHASE 2: Restricted Mandate (With closed canonical blockers)

2 rejections in 37 audits. Blockers strictly restricted to objective contract violations.



Model intelligence remained identical. What changed was the precision of the governance contract.

HOW TO GOVERN THE NORMATIVE GUARDIAN:

Strict prevention focus models require bounded jurisdiction. Explicitly define what constitutes a valid rejection before expecting auditor consistency. Without closed taxonomy, prudence becomes chronic blockage.

PREVENTION FOCUS SHADOW: ABSTINENCE WHEN CONTEXT IS PARTIAL

A new version is a new peer.
Opus 5.2 (30.8%) vs Opus 5.5 (57.1%).

We upgraded the code audit API from Claude Opus 5.2 to 5.5. A more capable Guardian tolerates zero ambiguity: it paralyzes and refuses to judge.

CLAUDE OPUS 5.2

Production Code QA

30.8%

8 rejections across 26 audits

Audited the strict code package provided with the task, tolerating minor file omissions and keeping engineering deliveries flowing.

CLAUDE OPUS 5.5 (ROUND 1)

Production Code QA

57.1%

20 rejections across 35 audits

Refused to evaluate in over 30% of calls demanding entire peripheral codebase, falling into the shadow of operational abstinence.

EXECUTIVE TAKEAWAY FOR VERSION UPGRADES:

When contextual prudence operates without exhaustive packages, it converts into refusal. Opus 5.5 demands stable evidentiary framing before assuming liability.

Golden Rule: Never upgrade model versions in production without recalibrating contracts.

COMBINING METACOGNITION AND PREVENTION IN THE SAME PIPELINE

Finding defects
is not verifying fixes.

When an auditor rejects code and dev patches it, preventive models tend to invent new requirements on return. We built the Anti-Goalpost circuit breaker:

01

DELIVERY OF CODE & EVIDENCE

Developer submits implementation accompanied by automated tests and physical PostgreSQL proofs.

02

ROUND 1 · GPT-6 ASTRA (THE METACOGNITIVE ARBITRATOR)

Fast calibrated audit (23.2s). Uncovers all real contract blockers without contextual abstinence.

03

SURGICAL & BOUNDED PATCHING

Engineering corrects strictly the items flagged in the official Round 1 rejection report.

04

ROUND 2 · CLAUDE OPUS 5.5 (GUARDIAN UNDER ANTI-GOALPOST)

Verifies exclusively whether Round 1 blockers were resolved. Forbidden from moving goalposts.

PRODUCTION RESULT AFTER ROLE INVERSION:

44 approvals across 44 Round 2 audits by Claude Opus 5.5.

Opus 5.5 did not become lenient: it received corrected scope and closed jurisdiction. Placing Arbitrator in discovery (R1) and Guardian in closure (R2) eliminated infinite loops.

PROFILE 2 · SURGICAL SPEED WITH METACOGNITIVE MONITORING (FLAVELL)

Fast does not mean permissive.
GPT-6 Astra: 7.8s & 20 escalations.

GPT-6 Astra was our most dispatched engine. We audited whether its high velocity compromised its ability to monitor its own limits:

MEDIAN DURATION PER AUDIT DISPATCH:

Pipeline Stage	GPT-6 Astra (Arbitrator)	Claude Opus 5.5 (Guardian)
Round 1 (Discovery)	**23.2 seconds**	**50.9 seconds**
Round 2 (Anti-Goalpost)	**7.8 seconds**	**24.8 seconds**

METACOGNITIVE CALIBRATION UNDER UNCERTAINTY (55 R2 AUDITS):

35
Direct Approvals (Complete Evidence)

20
Human Escalations (Declared Uncertainty)

0
Hasty Rejections without Evidence

VIRTUE OF THE METACOGNITIVE ARBITRATOR: KNOWING WHAT IT DOES NOT KNOW

Under Flavell's theory, metacognition is distinguishing confirmed fact from real uncertainty. In 36% of second-round audits, Astra recognized evidentiary limits and handed decision authority back to humans instead of guessing.

PREVENTION FOCUS (SONNET) VS. PROMOTION FOCUS (CODEX) REVIEWED BY SOL 6.1

Frontend and backend cast opposite shadows.

We deployed GPT-6.1 Sol as peer reviewer across 100 deliveries (50 frontend by Sonnet 5.5 and 50 backend by Codex 5.3). Each regulatory focus fails differently:

ANTHROPIC SONNET 5.5 (FRONTEND)

Shadow: Omission in Open Scope

50%

Incompleteness or lost context

- Declares lack of context in 57% of cases
- Loading and error states omitted
- Visual token degradation on long screens
- Accepted 100% of Sol 6.1 corrections

88 sessions · Median: 139s (open visual scope)

OPENAI CODEX 5.3 (BACKEND)

Shadow: Improvised Rules

30%

Silent assumption of business logic

- 0% context lack declarations
- Infers business rules when spec is silent
- Fills API gaps to close deliveries
- Accepted 90% of Sol 6.1 corrections

77 sessions · Median: 48s (promotion focus)

HOW PEER REVIEWER SOL 6.1 REINS IN BOTH (95% AVERAGE ACCEPTANCE):

For Sonnet on Frontend, the brake is providing visual context and demanding explicit state checklists. For Codex on Backend, the brake is tightening acceptance criteria in specifications before writing code.

PROFILES 3 & 4 · HIGH OPENNESS (OCEAN) VS. LOW AGREEABLENESS (JANIS)

Risk requires institutional friction.

According to Irving Janis, cooperative committees collapse into groupthink. Across 31 FMEA risk analyses, we paired the Explorer with the Socratic Adversary:

> *"S=10 with O=1 is faith, not engineering."*
>
> Rule coined by Grok 4.7 in our Second Brain to block ratings where a catastrophic severity (S=10) receives minimal occurrence (O=1) without physical evidence of controls.

THE DIALECTIC DYNAMIC BETWEEN EXPLORER AND ADVERSARY:

DEEPSEEK V4 PRO

The Exhaustive Explorer (Openness)

Light: Structural exhaustiveness in failure modes. Isolated Shadow: Catastrophizing or mitigation optimism (assigning low O and D to merely imagined defenses).

GROK 4.7 (xAI)

The Socratic Adversary (Anti-Consensus)

Light: Epistemological skepticism demanding verifiable tests and telemetry. Isolated Shadow: Performative skepticism if operating without final closing arbitration.

TAKEAWAY: RELIABILITY SPRINGS FROM INSTITUTIONALIZED CONTRADICTION

Whoever imagines the mitigation tends to believe in it. Final risk scores only achieved credibility after the Explorer's inventory survived the Socratic Adversary's scrutiny.

PROFILE 5 · MULTIMODAL EXECUTIVE CONTROL (BADDELEY) & JUNGIAN SHADOW

Executive Cortex:
Integration vs. Hyperagency.

Gemini 4 Argon High acts as Cognitive Maestro across engineering pipelines, connecting 5 layers of operational reality under governance:

LIGHT · VERIFIABLE ORCHESTRATION

5 Integrated Operational Layers:

- 1. Declared code & contracts (.mdoc)
- 2. Physical PostgreSQL (real SELECT)
- 3. Live application & network (F12)
- 4. TDD & mechanical gates (Track A)
- 5. Epistemic memory via MCP (Track B)
- Living manuals: -92% token consumption

SHADOW · AGENCY INFLATION (JUNG)

Hyperagency & Self-Sufficiency:

- Holding all tools tempts the model to resolve the entire delivery cycle alone.
- Risk of self-referential closure: interpreting rules, coding, creating tests, and self-approving.
- Illusion of sufficient integration: operational fluency cannot replace independent review.

CONSTITUTIONAL BRAKE: SEPARATION OF POWERS & BUSINESS SOVEREIGNTY

Whoever executes the delivery never audits their own work (Track A vs. Track B).

Just as we barred 81 blockers where auditors used privacy pretexts to legislate instead of the Product Owner, the Cognitive Maestro is strictly governed by separation of powers:

- **Cognitive Maestro (Argon): Integrates and executes Track A, but cannot self-approve.**
- **Executive Committee (Track B): Audits contracts, code, and risks with full independence.**
- **Human Product Owner: Holds exclusive sovereign authority over business and product logic.**

TAKEAWAY: The maturity of the Maestro lies in knowing when to summon the other 4 archetypes.

COGNITIVE ECOLOGY FOR ENGINEERING LEADERS AND ARCHITECTS

Do not hire a score.
Govern archetypes & shadows.

The highest engineering performance emerges when all 5 families operate under constitutional brakes that constrain each shadow:

01 **The Normative Guardian (Anthropic)**

Prevention Focus: requires closed taxonomy, full codebase, and Anti-Goalpost in R2.

02 **The Metacognitive Arbitrator (OpenAI)**

Metacognition: Astra in fast triage; Sol 6.1 guarding against Codex 5.3 improvised logic.

03 **The Exhaustive Explorer (DeepSeek)**

Broad Openness: maps failure modes across 31 FMEAs, but never scores risk in isolation.

04 **The Socratic Adversary (xAI Grok)**

Anti-Groupthink: challenges cosmetic scores and demands physical proof of controls in R2.

05 **The Cognitive Maestro (Google Argon)**

Multimodal Cortex: integrates code, DB, app, and TDD, without ever auditing its own work.

NEXT STUDY (POST 3) · PSYCHOLOGICAL THEORY AND MODEL DISSECTION:

Soon we will publish the full research on Higgins, Flavell, OCEAN, and Jung in production.

PUBLIC MCP SERVER FOR ARCHITECTS AND AI AGENTS:

`https://mcp.lb3.digital/sse` Inspect our live rules and telemetry.

FOR CTOS, FOUNDERS, AND ENGINEERING LEADERSHIP:

Which of these 5 cognitive shadows appears most in your engineering pipelines today?