NON-NAÏVE HYBRID HUMAN–AI EXECUTION PLAYBOOK

Version 1.0

Purpose

This Playbook defines the **operational discipline** required to implement the Hybrid Human–Agent Operating Standard in real systems without creating safety, ethical, or accountability failures.

It is mandatory for any deployment where Hybrid Human–AI outputs influence real-world decisions, actions, or access.

---

1. EXECUTION PRINCIPLE

> **AI may participate in cognition.**
> **Humans retain epistemic authority, moral responsibility, and outcome accountability.**

Execution exists to enforce this principle under pressure.

---

2. MANDATORY HUMAN ROLES

Every Hybrid Human–AI system MUST explicitly assign the following roles.

2.1 Problem Owner (PO)

- Defines the problem and non-goals
- Sets success and failure criteria
- Owns problem framing errors

2.2 Decision Owner (DO)

- Makes final decisions
- Accepts or rejects AI-influenced outputs
- Is accountable for downstream outcomes

2.3 Validation Lead (VL)

- Designs validation protocols
- Defines acceptable error
- Has authority to block deployment

If these roles are unclear, execution must stop.

---

3. PERMITTED AI ROLES

AI roles MUST be explicitly declared per task.

Allowed

- Hypothesis generation
- Option enumeration
- Counterfactual exploration
- Red-team critique
- Drafting and synthesis
- Pattern identification

Prohibited

- Problem definition
- Final decision making
- Risk ownership
- Ethical arbitration
- Accountability transfer

If AI implicitly assumes a prohibited role, the system is misconfigured.

---

4. PHASE-GATED EXECUTION WORKFLOW

No phase may be skipped.

Phase 1 — Problem Framing (Human-Only Gate)

Required artifacts:

- Problem definition (1 page)
- Explicit non-goals
- Ethical and legal constraints
- Known unacceptable failure modes

AI may critique *after* human framing is complete.

Phase 2 — AI-Augmented Exploration

AI is used to:

- Expand solution space
- Identify edge cases
- Stress assumptions
- Surface alternative approaches

Outputs are **options**, not recommendations.

Phase 3 — Human Narrowing & Judgment

Humans must:

- Select candidate approaches
- Reject others explicitly
- Identify AI weaknesses

Mandatory question:

> If this fails, how does it fail, and who is harmed first?

Phase 4 — Validation Design (Pre-Deployment)

Validation MUST be designed before execution.

At least one of:

- Expert review
- Simulation
- Shadow mode
- Historical back-testing

- Limited pilot

Define:

- Acceptable error
- Kill-switch criteria
- Escalation paths

If validation cannot be defined, restrict AI to advisory-only use.

Phase 5 — Controlled Deployment

Requirements:

- Human decision authority with veto power
- Full logging of prompts, outputs, overrides
- No silent automation
- Clear documentation of AI contribution

Forbidden language:

- “The AI decided…”
- “We followed the model…”

Phase 6 — Monitoring & Feedback

Continuously track:

- Errors and near-misses
- Human override rates
- Automation bias indicators
- Context drift

Regular reviews are mandatory.

---

5. TRUST CALIBRATION RULES

- High trust: pattern generation, drafting
- Low trust: factual claims, inference
- Zero trust: safety-critical decisions

Fluency does not equal correctness.

---

6. ACCOUNTABILITY STATEMENT

Every deployment MUST include:

> “The accountable human for decisions influenced by this system is:
> Name, role, escalation path.”

No accountability → no deployment.

---

7. STOP CONDITIONS

Execution MUST pause if:

- Error thresholds are exceeded
- Context materially changes
- Human override rates spike
- Outputs become unexplainable

---

8. WHAT THIS PLAYBOOK OPTIMIZES FOR

Not speed.
Not scale.
Not convenience.

But:

- Bounded risk
- Traceable decisions
- Durable trust
- Safe iteration

FINAL WARNING

Most failures in Hybrid Human–AI systems are **execution failures**, not model failures.

This Playbook exists to prevent human abdication disguised as automation.