NON-NAIVE HYBRID HUMAN-AI EXECUTION PLAYBOOK

Version 1.0

How This Playbook Is Intended to Be Used

> ***This section defines when, how, and by whom this Playbook must be applied.***

Requirement	Description
When it is mandatory	Any deployment where Hybrid Human-AI outputs influence real-world decisions, actions, or access.
Who must own it	A named human accountable for execution - not a team, not a committee.
What happens if it is ignored	Deployment proceeds without governance coverage. Accountability becomes undefined. Risk is unbounded.
How it relates to the Standard	The Hybrid Human-Agent Operating Standard defines governance constraints. This Playbook defines how to execute them. They are designed to be used together.

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

> ***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

> ***NON-NEGOTIABLE***
>
> *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

> ***WARNING***
>
> *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)

Gate	Requirement
ENTRY CONDITIONS	Problem Owner identified. Scope defined.
EXIT CONDITIONS	All required artifacts complete. Human sign-off obtained.

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

Gate	Requirement
ENTRY CONDITIONS	Phase 1 complete. Problem framing approved.
EXIT CONDITIONS	Options documented. Human review scheduled.

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

Gate	Requirement
ENTRY CONDITIONS	Phase 2 complete. Options documented.
EXIT CONDITIONS	Candidate approaches selected. Rejections documented.

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)

Gate	Requirement
ENTRY CONDITIONS	Phase 3 complete. Candidates selected.
EXIT CONDITIONS	Validation protocol approved. Kill-switch defined.

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

> ***NON-NEGOTIABLE***
>
> *If validation cannot be defined, restrict AI to advisory-only use.*

Phase 5 - Controlled Deployment

Gate	Requirement
ENTRY CONDITIONS	Phase 4 complete. Validation passed.
EXIT CONDITIONS	Deployment live. Monitoring active.

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

Gate	Requirement
ENTRY CONDITIONS	Phase 5 complete. System deployed.
EXIT CONDITIONS	Ongoing. Regular reviews scheduled.

Continuously track:

- Errors and near-misses

- Human override rates
- Automation bias indicators
- Context drift

Regular reviews are mandatory.

---

5. TRUST CALIBRATION RULES

> ***Any deviation from this trust calibration requires written justification.***

Trust Level	AI Role	Human Requirement
High trust	Pattern generation, drafting	Human review before use
Low trust	Factual claims, inference	Human verification required
Zero trust	Safety-critical decisions	AI advisory only, human decides

Fluency does not equal correctness.

---

6. ACCOUNTABILITY STATEMENT

> ***ACCOUNTABILITY DECLARATION***
>
> *Every deployment MUST include the following statement, completed and signed:*
>
> ---
>
> *“The accountable human for decisions influenced by this system is:*
>
> ***Name:*** *_____________________*
>
> ***Role:*** *_____________________*
>
> ***Escalation Path:*** *_____________________”*
>
> ---
>
> *This statement must be copied verbatim into deployment documentation.*

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

> ***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.*

APPENDIX A - EXECUTION ANTI-PATTERNS

The following patterns represent common compliance failures. Their presence indicates governance breakdown, not compliant implementation.

Anti-Pattern	Why It Fails
“Human-in-the-loop” without veto power	Human presence without authority is theater, not governance.
One-shot prompting used as decision support	No validation, no iteration, no accountability trail.
Validation deferred to production	Risk transferred to end users without consent.
Accountability assigned to a team, not a person	Diffused accountability is no accountability.
“The AI recommended it” as justification	AI cannot bear responsibility. This is abdication.
Skipping phases “because we’re experienced”	Experience does not exempt systems from governance.

> ***These anti-patterns prevent plausible deniability.***
>
> *No one can claim “we thought we were compliant” if these patterns are present.*

Non-Naive Hybrid Human-AI Execution Playbook v1.0 Companion to the Hybrid Human-Agent Operating Standard