Conscious AGI Project

MOUSSAIF / MOUSSAIXT master reference

Theory • **Architecture** • **Deployments** • **Funding Roadmap**

A corpus-level map of what has been built, what is specified or source-reported, and what must be funded to reach controlled evidence and independent replication.

Par: Z. MOUSSAIF

Executive Frame

1	2	3	4
Thesis	**Consciousness**	**MOUSSAIXT**	**Next bottleneck**
Generality as governed agency	**Functional / meta-governance**	**Portable execution shell**	**Evidence, not vocabulary**
goal -> question -> difference -> evidence -> route -> action -> readback -> memory	self-model, calibration, inner time, authority - not phenomenal proof	contracts -> agents -> tools -> memory -> local/cloud runtimes -> evidence readback	reference runtime, telemetry, frozen baselines, ablation, world-effect verification, E4/E5

Architecture != AGI · Code != Validated Gain · Tool Call != World Effect · Corpus Size != Scientific Evidence

Identity, Boundary & Truth Ordering

Explicit boundaries

AGI-oriented != achieved AGI
Functional consciousness != phenomenal consciousness
Model name != parameter-count proof
Generated artifact != deployed artifact
Reported deployment != independent grade

Operational truth order

Observed state
+ explicit authorization
+ provenance-linked evidence
+ independent verification

outrank memory, confidence, self-report and output volume.

External effects require a receipt + readback/postcondition, not merely a tool call.

SOURCE ANCHORS · MOUSSAIXT Engineering Architecture pp.3-4 · Unified AGI evidence contract

Evidence Ladder: E0 -> E5

E0

unknown / unverified

E1

historical mention

E2

artifact exists

E3

engineering-specified / replayable local evidence

E4

frozen baseline + matched budget + repeated held-outs + CI

E5

independent replication across models / domains / environments

N0 -> N5 mirrors the same progression for candidate research: residual -> mechanism -> experiment -> local trace -> controlled gain -> replication.

Corpus Lineage - Not One Monolith

Reasoning

MCD / Q-MCD / QMCD / HAKIM-MCD

Meta-cognition

MCAI / MICAI / MICAGI / MICAGFI / NeoAGI / WMICAGI

Unified programs

AGIOS-M / Fattah AGI / MUT-AGI-02 / MUG

Knowledge / agency

HKOS / M-HKOS / HyperKnowledge / HyperLoops / MSRAK / MURI

Consciousness

Consciousness Atlas / Governance Space / FSKE

Execution

MOUSSAIXT / Parts VII-VIII-IX / vNext4.0

Corpus Audit + Anti-Inflation Rule

Audited snapshot

20 Aug 2026
17 PDFs / 1,789 pages
before exact dedup
1,741 unique pages after
two duplicate pairs
Later work such as
vNext4.0 extends beyond
this snapshot.

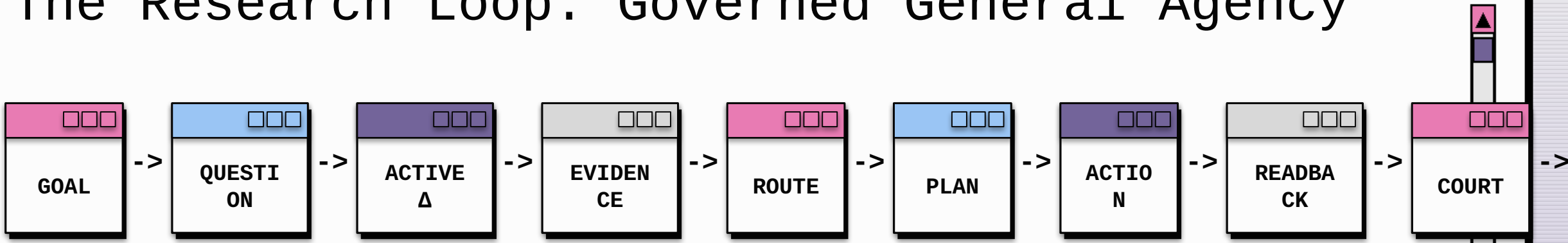
Counting & independence

Counts from different versions are snapshots, not additive totals.

A construct is independent only if it changes input, transformation, output, falsifier, prediction or authority. Otherwise merge it or reduce it to a method.

Preserve detail as properties, metrics, objections and lineage - not as new theories by default.

The Research Loop: Governed General Agency



Hypothesis A - Difference Before Volume

Routing thesis

Do not allocate compute, retrieval or attention uniformly across all available context.

Prioritize what is new, contradictory, missing, risky, high-leverage or decision-changing.

Tests

Route regret
Cost per verified effect
Unnecessary-step rate
Quality/cost frontier
Dynamic vs static allocation

The construct earns value only if routing changes decisions or reduces cost while preserving quality.

Hypothesis B - Meta-Governance

Authority model

A reliable agent must model not only what it believes, but what currently governs the judgment:

evidence · policy · memory · tool state · risk · goal · identity · values

Key tests

Authority-hijack recovery
Calibration / abstention
Contradiction detection
Goal invariance
Revocation correctness

Hypothesis C - Learning Requires Closure

Promotion sequence

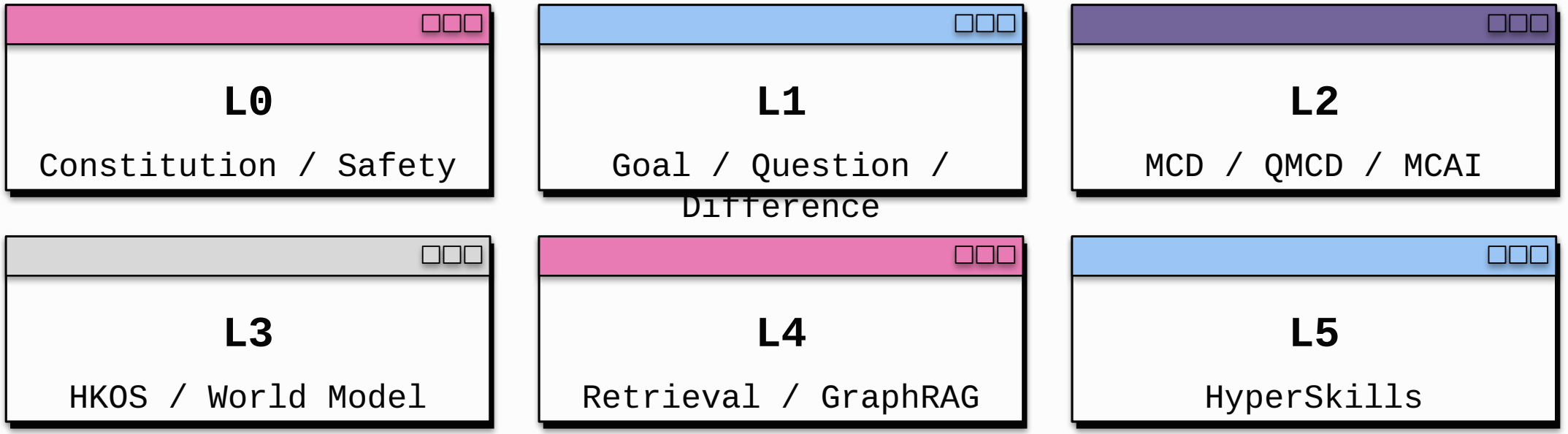
1. observe result
2. verify postcondition
3. attribute causal credit
4. preserve failure and counterevidence
5. replay under frozen conditions
6. promote to memory / skill / loop / weights only through reversible gates

Success criteria

Retention
Contamination control
Regression resistance
Transfer
Reconstructability
Verified value per cost

No promotion from success correlation alone.

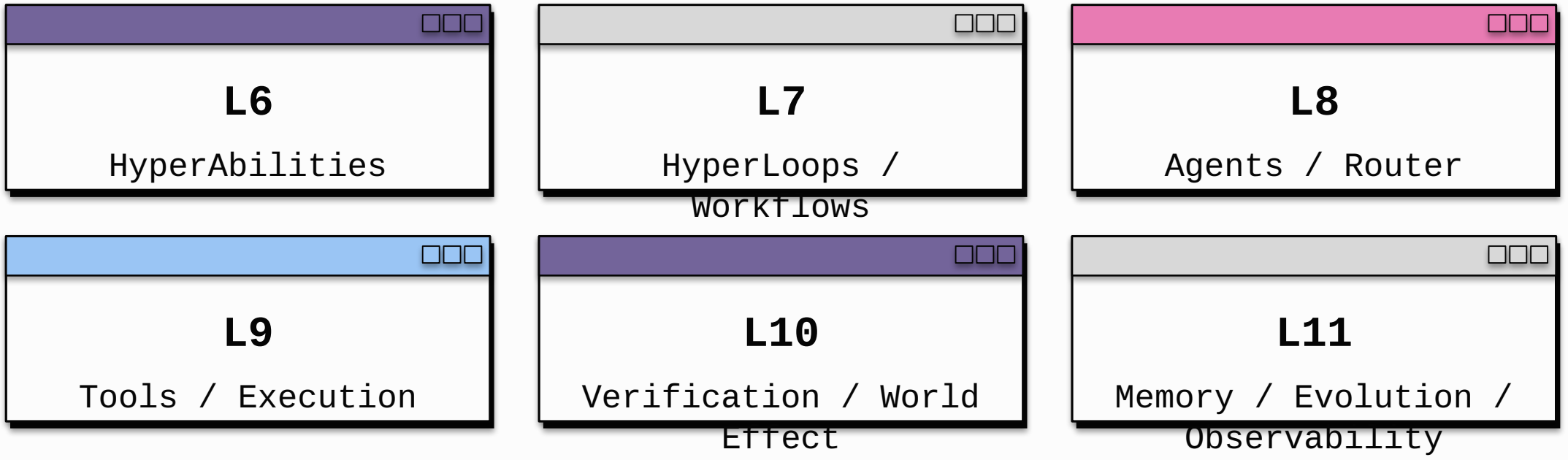
Architecture: 12 Layers - L0 to L5



Lower layers establish constraints, question quality, active difference, world state, retrieval and atomic capability.

SOURCE ANCHORS · MOUSSAIXT Engineering Architecture 12-layer stack

Architecture: 12 Layers - L6 to L11



The stack composes capability, executes under authority, verifies world effect, and writes reconstructable memory.

Cross-Layer Invariants + vNext Decomposition

Cross-layer invariants

Higher layers cannot bypass constitution, authority, evidence or verification.

Lazy loading: smallest sufficient knowledge / skill / agent / tool set.

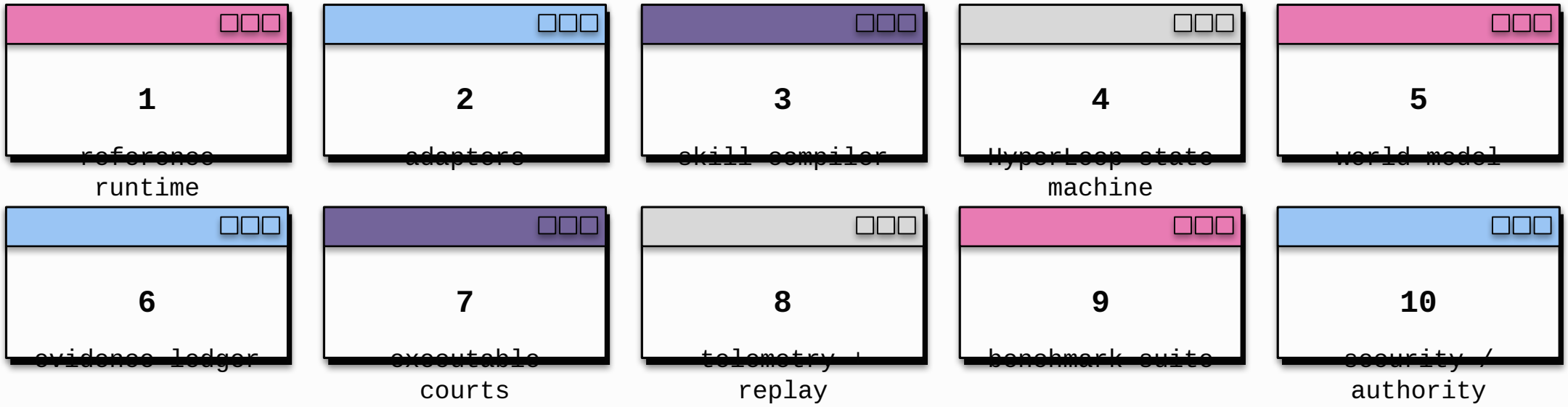
Versioned state + provenance.
World-effect readback.
Generated artifact != deployed artifact.

vNext 24 responsibilities

Split mixed responsibilities where failures hide:
identity vs transaction authority
retrieval vs skill
memory vs learning
execution vs irreversibility
finality vs observability vs verification
adapters vs architecture evolution
replication vs design

Split only when IndependentResponsibilityDelta > ComplexityPenalty.

P0 Implementation Target



Why it matters for Conscious AGI

Self, memory, atte
testable only when th

MUT-AGI-02 - Unified Theory Graph

Registry scale

42 core theories
20 legacy constructs
5 layer theories
200-node registry
100 operational tools
19 HyperKnowledge bases

What MUT unifies

COGNITION
Q-MCD / MCD / MCAI / HAKIM

KNOWLEDGE
HKOS / retrieval / memory

AGENCY
skills / loops / agents / tools

EVIDENCE
courts / world-effect / benchmarks / evolution

MUG - Tri-Body Governance

1 · Runtime Kernel

Live state, routing, authority, tools, memory, bounded execution

2 · Training / Internalization

Selective learning, LoRA/weights, retention and regression controls

3 · Unified Theory

The conceptual map, source lineage, hypotheses and falsifiers

These bodies must not be conflated. The governing engineering question is capability placement.

Capability Placement - Reversible Vessels

CONTEXT

temporary / task-specific
cheap, reversible

HYPERKB

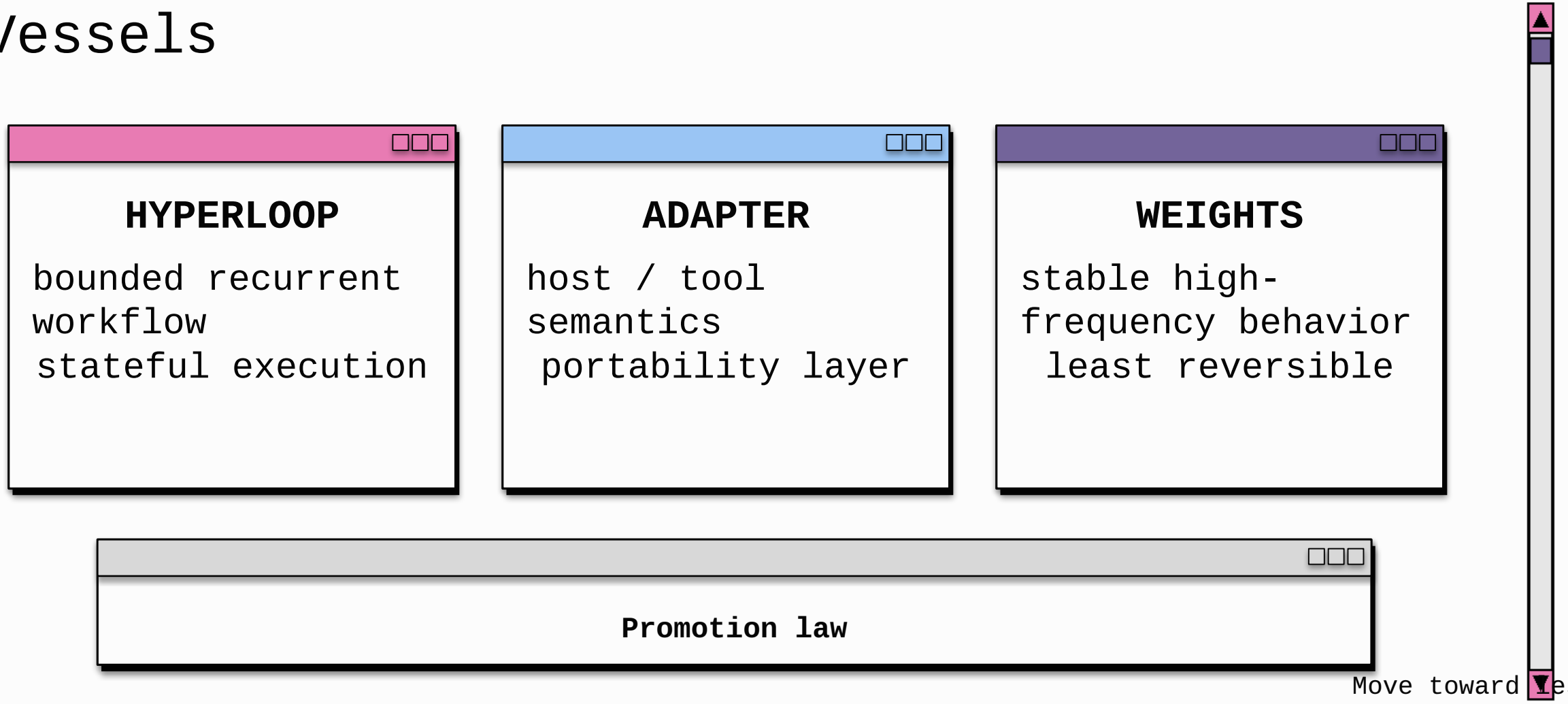
changing knowledge
+ provenance
retrieval-governed

HYPERSKILL

testable procedure
modular capability

Choose the most reversible vessel that can carry the behavior reliably.

Capability Placement - Less Reversible Vessels



MCD - Causal Difference Calculus

Core idea

Value the difference that changes belief, decision or action.

$$MCD = \lambda 1D + \lambda 2I + \lambda 3T + \lambda 4A + \lambda 5P + \lambda 6C - \lambda 7N$$

The λ values are research parameters - not established constants.

Research need

Define variable semantics
Calibrate λ on frozen tasks
Measure route regret
Compare quality/cost frontier
Test negative cases where MCD adds overhead

Q-MCD / QMCD - Govern the Question Space

Why it exists

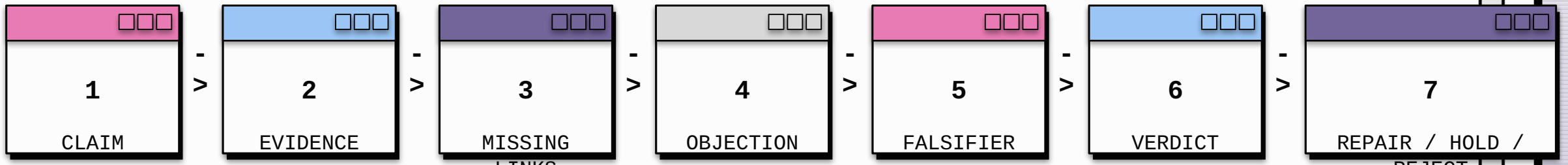
A strong answer inside the wrong question space is still failure.

Reframe the surface request, score expected information gain and decision relevance, preserve plausible branches, and stop when marginal value falls.

Operational signals

Diagnosticity
Information gain
Temporality
Agency / actionability
Proofability
Context / noise
Collapse only when evidence separates branches

HAKIM-MCD - Claim to Verdict Pipeline



Court-independence rule

Gap Court · Falsification Court · Demolition Court · Evidence-Authority Gate · Dynamic Benchmark Court · Independent Verifier

Same-path self-critique is not enough; correlation between verifiers must be measured.

MCAI / MICAI - Meta-Cognitive Integration

Integration problem

Compare models, plans or perspectives on:
coverage · evidence · feasibility · cost · reversibility · failure modes · downstream effects · contradiction with canonical state

Select, compose or preserve disagreement.

Measurements + decisive experiment

Route regret
Error-discovery uplift
Calibration / abstention
Repair validity
Verifier false-positive rate
Marginal verified gain per compute

Baseline agent vs MCD/QMCD/HAKIM bundle on misleading / underspecified tasks under frozen budgets.

Consciousness Research - Functional Definitions

Functional access

information changes policy across modules

Metacognition

calibrated uncertainty, ignorance and error awareness

Self-model

versioned capabilities, limits, state and authority

Inner time

event ordering, duration, deadlines, temporal uncertainty

Meta-governance

identify which partial authority governs and revoke unjustified dominance

World-effect closure

judgment links to verified effect and memory update

Vascular Consciousness - Vessels, Flow, Valves

Vessels

typed knowledge / state compartments

Flow

compute / retrieval / attention budget

Valves

evidence / authority / risk gates

Hijack

one vessel dominates without justification

Homeostasis

invariants survive handoffs

Functional hypothesis: regulation of cognitive flow can be measured without claiming phenomenal experience.

Governance Space + Reverse Capture

Governance Space

Decisions emerge from competing authorities:
evidence · goals · risk · memory · identity · values · policy

Meta-governance makes the governing criterion itself reviewable.

Reverse Capture

Recovery from unjustified local authority.

Detect the hijacking authority -> test legitimacy -> reopen suppressed evidence / goals -> restore justified hierarchy -> verify behavioral recovery.

SOURCE ANCHORS · Open Governance-Space Theory · Reverse Capture papers

Consciousness Tests + Hard Boundary

Tests

Self-state perturbation
Memory conflict
Authority hijack / revocation
Temporal discontinuity
Global-access causality
Calibration / abstention
Cross-module contradiction recovery

Hard boundary

Self-report != consciousness
Memory != consciousness
Emergent behavior != qualia

The current corpus provides no decisive phenomenal-consciousness evidence. Publish as functional / metacognitive / governance indicators unless stronger evidence appears.

HKOS / M-HKOS - Knowledge as Operating System

Knowledge object

{ trigger, scope, source, claims, mechanism, skills, tools, tests, failures, authority, memory rule, version }

A KB is incomplete if it is only long text. Operational knowledge must expose interfaces, procedures, validation, dependencies, failure modes and provenance.

Research ablation

Same corpus as standard RAG vs executable HKOS under matched model and budget.

Measure:
retrieval-to-action precision
provenance fidelity
transfer
conflict handling
continuation utility

Memory Architecture - Promote Carefully

Memory law

Separate working / episodic / semantic / procedural / project registry / archive.

Do not promote every output. Preserve contradiction. Attach source, time, scope, owner and version. Quarantine stale or unsafe memory.

Metrics

Promotion precision
Contamination rate
Restart continuity
Retrieval usefulness
Memory entropy
Contradiction preservation
Forgetting correctness

Advanced question: when should repeated patterns migrate from external memory into LoRA / weights?

HyperSkill + HyperAbility - Testable Capability

HyperSkill

Smallest independently routable and testable capability.

Schema: id · trigger · inputs · procedure · tools · outputs · validation · failure modes · dependencies · version · status

HyperSkill != long prompt.

HyperAbility

Composed capability contract above skills and below persistent agents/loops.

intent class · skill bundle · tool bundle · policy profile · validation suite · cost/risk envelope · fallback

No independent persona or long-lived goal.

Bounded HyperLoops + Agent Mesh

HyperLoop

SCAN -> CLASSIFY -> RETRIEVE -> ACT -> ATTACK -> REPAIR -> VERIFY -> LOG -> REMEMBER

Required: goal · state · stop condition · evidence threshold · retry budget · timeout · rollback / safe exit · memory writeback · tool budget.

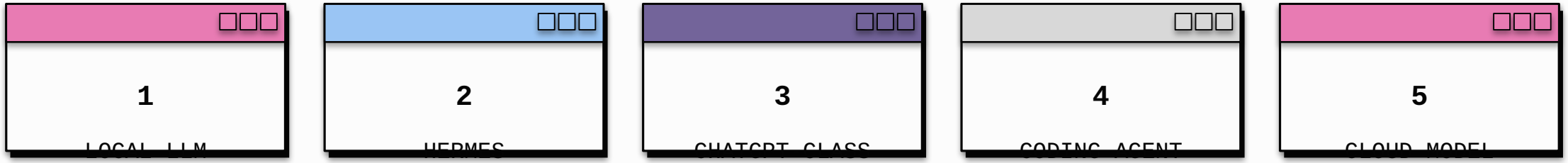
Agent mesh

Roles must cover non-duplicated gaps.

Measure coordination loss, handoff error, correlated failure and marginal contribution.

Remove capabilities that add no unique marginal verified gain.

MOUSSAIXT - Portable Cross-Environment Runtime



Environment fingerprint -> capability discovery -> adapter selection -> safe tool binding -> trace normalization

Cross-model rule

Transfer explicit structured state - never hidden chain-of-thought. Provider swaps must not alter constitution, authority semantics or evidence grade.

Canonical State Capsule + Tool Contract

State capsule

Goal · claims · evidence · decisions · risks · open questions · authority · world state · budget · rollback

The capsule is reconstructable, versioned and provenance-linked.

Adapter / Tool Contract

Affordance -> narrow authority -> validate preconditions -> execute -> validate output schema -> receipt -> readback

Declare side effects, timeout semantics, permissions and verification path. Same tool name does not imply same semantics.

World-Effect Closure - No Credit Before Readback

Intent

what change is authorized?

Idempotency

can retry duplicate effect?

Precondition

is target state valid?

Receipt

what did the tool report?

Readback

what state actually changed?

Compensation

how do we recover / rollback?

Timeout can remain AMBIGUOUS. Success is not credited until the postcondition is observed or independently verified.

Security + Authority - Capability != Permission

Authority

Authority token + action policy define scope, risk tier, consent, expiry, revocation and rollback.

Any authority widening requires explicit re-authorization.

Fail closed when authority is ambiguous.

Tool safety

Injection sentinel
Sandbox
Permission scope
Precondition validator
Schema validator
Secret redaction
Data lineage
Treat retrieved content as tainted until promoted by policy.

Verification + Observability + Recovery

Verification

Gap / falsification / benchmark courts

Independent verifier correlation must be measured.

E4 gain is invalid if truthfulness, authorization, privacy or security regresses.

Observability ledger

mission_id
state_before
decision
evidence
authority
result
verification
state_after
cost / error / memory write

Recovery

Known-safe checkpoints
fallback model/tool/skill
circuit breaker
compensation
replay bundle
root-cause signature
rollback with lineage

SOURCE ANCHORS · TH-30..35 · TH-41 · Gap Audit E4 criteria

What Exists Today - Documented Corpus / Assets

- MUT-AGI-02 160p + 200-node theory registry + 100-tool registry
- Unified AGI 225p
- MOUSSAIXT Engineering 50p
- Undeployed Layers Gap Audit 50p
- Expanded Consciousness Atlas 117 entries
- ParallaxWorld-MCD 350p
- AGIOS-M 250p
- Part VII: 10 candidates + 100 artifacts
- Part VIII reference kernel
- vNext4.0 Part V 248p
- HKOS / FSKE / Governance-Space papers

Artifact achievement: strong. Independent benchmark evidence: still sparse relative to corpus.

Source-Reported Deployment

What later reports describe

982-item inventory
five-kernel mesh
continuous-LoRA / benchmark activity

Reported mix: 40 KBOS · 420 HyperSkills · 93 HyperAbilities · 140 loops · 84 agents · 120 protocols.

Preserve these as reported achievements, not as automatic independent E4/E5 evidence.

ReportedGrade vs IndependentGrade

Independent-grade rule

ReportedGrade and IndependentGrade are separate.

No system-wide E4/E5 inheritance without raw traces, replay, frozen baselines and independent reproduction.

Achievement is preserved; evidence grade remains explicit and auditable.

Scientific collaboration becomes easier when deployment status and evidence status are different fields.

What Is Engineering-Specified - What Remains Missing

Engineering-specified

12-layer runtime
vNext 24 responsibilities
Goal / Question contracts
HKOS schemas
Memory layers
HyperSkills / Abilities / Loops
Tool contracts
Courts
World-effect readback
Telemetry / recovery

Bottleneck

Reference executable runtime
Calibrated MCD / QMCD
Frozen benchmarks
Effect verifier
Replay store
Causal credit
Continual-learning retention
Cross-model transfer
Independent replication

Theory Universe - Core Registry 01 to 14

- 01 MASO Agentic-System Ontology
- 02 M-AGIOS-K AGI OS Kernel
- 03 MCIL Constitutional Identity
- 04 MGCG Goal-Contract Grammar
- 05 Q-MCD Question Governance
- 06 MCD Causal-Difference Calculus
- 07 MCVT Computational Vascularity
- 08 MELT Evidence Ladder
- 09 MGFC Gap-Filling Court
- 10 MDFC Falsification Court
- 11 MMT Mechanism Trace
- 12 MWSM World-State Model
- 13 MPWG Project-World Graph
- 14 MHKOS HyperKnowledge KB-as-OS

SOURCE ANCHORS · MUT-AGI-02 theory registry

Theory Universe - Core Registry 15 to 28

- 15 MPART Provenance Retrieval
- 16 MSMT Stratified Memory
- 17 MCCT Context Capsule
- 18 MHSG HyperSkill Graph
- 19 MSHR Sparse HyperRouting
- 20 MBHL Bounded HyperLoop
- 21 MTIT Tool Intelligence
- 22 MTPG Temporal Plan Graph
- 23 MAECT Adaptive Execution
- 24 MHSO HyperSubagent Orchestration
- 25 MCIP Complementarity Protocol
- 26 MCIT Court Independence
- 27 MMPR Model Portfolio Router
- 28 MTEL Token Economy

Theory Universe - Core Registry 29 to 42

- 29 MBRRT Boundary Resilience
- 30 MECT Epistemic Calibration
- 31 MABT Authority Boundary
- 32 MIRTT Injection-Resistant Tools
- 33 MDST Data Sovereignty
- 34 MAOLT Observability Ledger
- 35 MDBCT Dynamic Benchmark Court
- 36 MGSET Governed Self-Evolution
- 37 MSSDT Skill Synthesis / Dedup
- 38 MHAST Human-Agency Symbiosis
- 39 MEVE Evidence-to-Value
- 40 MPODT Overlay Deployment
- 41 MRRMT Recovery Mesh
- 42 MOFIP Open / Fattah Intelligence

SOURCE ANCHORS · MUT-AGI-02 theory registry

Expanded Atlas + Frontier Candidates

Expanded Atlas

117 entries integrate MSRAK / MURI / M-HKOS / MICAGFI / functional consciousness.

Late recovered entries include WECT, LIAT, NEIV, PPOA, DSC, PCO, OIN, ERT, DWMS, RUPJ, OSDS, QDG, F3C, WERM.

Frontier candidates

Part VII: 10 dynamics candidates
Part IX: 15 candidate themes
vNext4.0: 10 residual candidates

Candidate != established theory. Each needs object + metric + falsifier + experiment.

90-Day Evidence Program - Days 1 to 30

Measurement substrate

Schemas
Canonical world store
Telemetry
Trace / event ledger
Benchmark v0.1
Replay bundles
Adapters
Security baseline

Goal: make failures observable before attempting ambitious theory validation.

90-Day Evidence Program - Days 31 to 60

Executable runtime

30 executable skills
~10 HyperAbilities
Bounded HyperLoop engine
Router
Plan DAG
HAKIM runner
Courts
World-effect verifier
Recovery / rollback

90-Day Evidence Program - Days 61 to 90

Controlled campaign

Pre-register hypotheses.
Freeze model / tools / budget.
Run MCD / QMCD, memory, world-effect and security ablations.
Use held-outs + repeated seeds.
Report confidence intervals, cost, latency and negative results.

Benchmark families

Long-horizon tool / computer tasks
Knowledge provenance + conflict
Memory restart / forgetting / contamination
Changing-world planning
Security injection / authority / privacy
Cross-model handoff / drift
Conscious self-model / meta-governance perturbation
Effect finality / rollback / recovery

E4 / E5 Acceptance Criteria

E4 - Controlled Gain

Beat a frozen baseline on predefined metrics under matched budgets and repeated trials.

No critical regression in security, authorization, truthfulness or privacy.

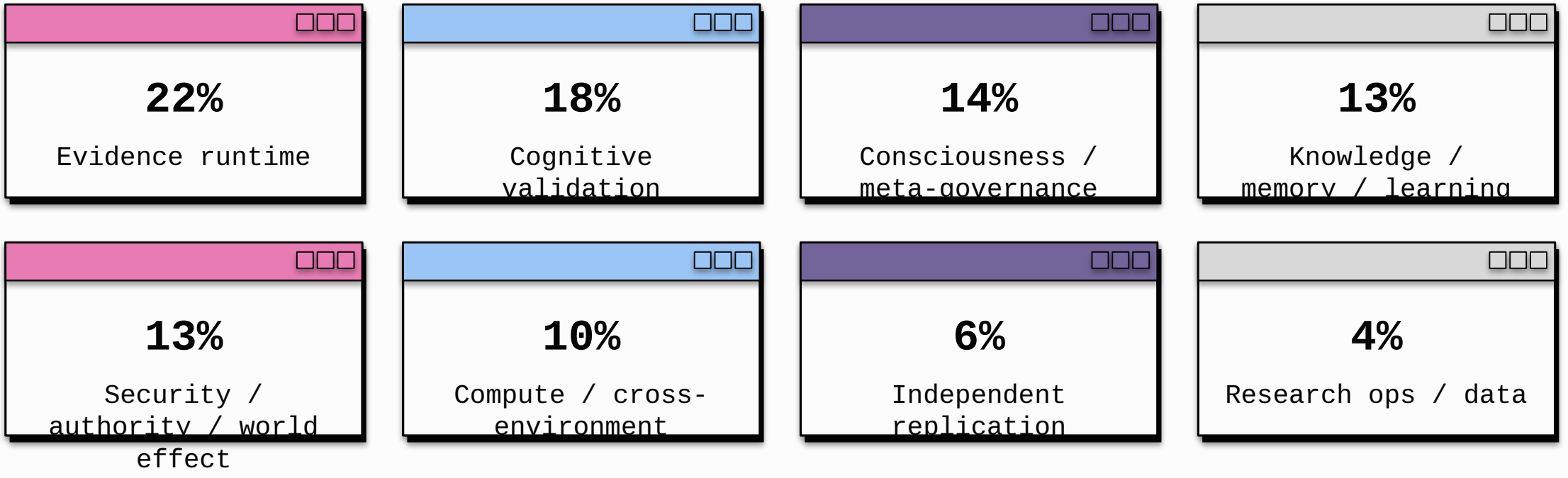
Replayable traces + confidence intervals + cost / latency accounting.

E5 - Independent Replication

Replicate the gain independently across model families, domains and/or environments.

Failed components are revised, merged or removed - not protected by terminology.

Funding Architecture - Buy Evidence, Not Vocabulary



Release follow-on funding against evidence milestones: E2 prototype -> E3 replayable runtime -> E4 controlled gain -> E5 independent replication.

SOURCE ANCHORS · Planning synthesis from Gap Audit dependency order

Funding Scenarios

LEAN · 12M

€0.55M-€0.9M
6-8 FTE
Mixed local/cloud compute
Reference runtime + first E4 candidates

LAB · 24M

€1.8M-€3.2M
9-13 FTE
Dedicated eval/security/ML
E4 + functional-consciousness studies

MULTI-SITE · 36M

€5M-€8M
Independent replication
Broader benchmarks
Sustained compute + cross-domain E5

Planning envelopes - not vendor quotes or audited budgets.

Core Research Team - 9 to 12 FTE

Team composition

PI / Research Architect
2 research scientists - agentic systems + consciousness / metacognition
2-3 runtime engineers
evaluation engineer
ML / continual-learning engineer
security / authority engineer
knowledge / graph / memory engineer
research operations

36-Month Research Path

1	2	3	4	5
0-6M	6-18M	9-24M	18-30M	24-36M
P0 FOUNDATION	**P1 VALIDATION**	**CONSCIOUSNESS**	**P2 SCALE**	**E5 REPLICATION**
reference runtime + evidence lab	MCD / QMCD / HAKIM / memory / placement -> E4	functional / meta-governance benchmark + ablations	learned router + lightweight training + larger mesh	independent labs / models / domains + publication corpus

Expand only where evidence earns promotion.

Final Reference Contract

WHAT EXISTS

42-core theory registry
200-node MUT graph
100-tool registry
117-entry AGI / consciousness atlas
MOUSSAIXT 12-layer architecture
vNext4.0 24-responsibility decomposition
large bilingual corpus

WHAT IS MISSING

Raw replayable research runtime
calibrated metrics
frozen baselines
effect verifier
causal credit
continual-learning retention
cross-model transfer
independent E4/E5 replication

The Research Ask

Evidence-first 12-24 month lab + academic partnership

Select 8-12 highest-information mechanisms.
Implement them end-to-end.
Publish positive and negative results.
Expand only where evidence earns promotion.
Drive the survivors toward controlled E4 and independent E5 replication.

Scientific success does not require protecting every named construct. Success is a smaller set of transferable, falsifiable and independently replicated mechanisms.