MOUSSAIF UNIFIED AGI · PART XIV · P14-B · 2026

الورقة الثانية: المسار الحامل للبرهان وهندسة الالتزام والسلطة

& Proof-Carrying Trajectories, Commitment Geometry Authority-Preserving Delegation

P14-B - إغلاق المسار والالتزام والسلطة

ضمن MOUSSAIF Unified AGI Part XIV - Executable Theoretical Completion

Moussaif Zouhair / زهير مساعف · MUT × MUG × MOUSSAIXT

الحالة العلمية: تركيب نظري/صوري/تصميم تجريبي. لا يمثل إثباتاً على تحقق AGI أو الوعي الظاهراتي، ولا يحول specification إلى evidence.

الملخص والأطروحة

تعالج هذه الورقة فجوة محددة في برنامج MOUSSAIF: نقل proof-carrying من action مفرد إلى trajectory كاملة، تمثيل هندسة الالتزام، وحفظ السلطة أثناء delegation عبر tools/agents/protocols. لا تقدم مساهمة الورقة على أنها «إثبات AGI»، بل كحزمة فرضيات ذات state variables وآليات قابلة للتعطيل وأدوات وتجارب ومفندات. معيار القيمة هو ما إذا أضافت النظرية فرقاً سببياً فوق baseline بسيط وتحمّل ذلك الفرق النقل بين شروط مختلفة.

تستوعب الورقة السلالة الداخلية عبر canonical owners بدل نسخ المحتوى: MURI، HKOS/M-HKOS، MICAI/MICAGI/MICAGFI، MCD/QMCD/MCAI/ Parts X-XIII، MOUSSAIXT، MUG، reconstructive memory، world-effect، MSRAK. المجتمع العلمي يوفر خطوط ضغط حديثة في long-horizon agents، control، multi-agent governance، benchmark unification، process reward models، memory، world models وself-improvement.

> **المساهمة:** نظرية مساعف للمسار الحامل للبرهان؛ نظرية مساعف لهندسة أفق الالتزام؛ نظرية مساعف لحفظ السلطة عبر التفويض. كل مساهمة مرتبطة بمعادلة design-level، feature flag، falsifier، تجربة transfer envelope، matched-budget، وأداة runtime.

سؤال الورقة: ما الحد الأدنى من البنية الإضافية التي تجعل إغلاق المسار والالتزام والسلطة قابلاً للقياس والتفنيد والربط بأثر العالم، دون تحويل density النظرية إلى claim غير مدعوم؟

العقد الإبستيمي ومنع تضخم الادعاء

طبقة	الوضع في الورقة
External literature	مرساة لمشكلة أو baseline أو mechanism prior؛ ليست دليلاً على صحة M14.
MOUSSAIF lineage	مصدر الهوية، التعريف، العلاقات، والمقترحات؛ evidence grade مستقل.
Equations	Formal specifications/heuristics حتى تتوافر estimators ومعايرة ونتائج.
Tools	SPECIFICATION إلى أن يوجد runtime trace وtests.
Experiments	تصاميم قابلة للتنفيذ؛ لا تُنسب نتائج غير منفذة.

$$\text{Claim strength} \leq \min(\text{Source strength}, \text{Measurement strength}, \text{Causal isolation}, \text{Replication strength})$$

كل نظرية تحصل على negative activation case، لأن المكوّن الذي «يفيد دائماً» يصعب التفريق فيه بين آلية حقيقية وبين تعقيد إضافي. وكل نجاح يُقارن بـnearest parent وbaseline بسيط وsham scaffold.

مصفوفة الضغط من حالة الفن 2026

المصادر التالية لا تُستخدم لزخرفة المراجع؛ كل مصدر يفرض قيد تصميم أو benchmark أو failure mode يجب أن يواجهه المسار.

E06 - Long-Horizon Agent Trajectory Attribution. arXiv:2608.06909. Fine-grained attribution benchmark with >1,300 annotated trajectories and leave-one-out baselines.

E09 - Aligning Agents via Planning: Plan-RewardBench. ACL 2026, 10.18653/v1/2026.acl-long.1062. Trajectory-level reward modeling; evaluator performance degrades with long trajectories.

E10 - ToolPRMBench: Evaluating and Advancing Process Reward Models for Tool-using Agents. Findings ACL 2026, 10.18653/v1/2026.findings-acl.602. Step-level PRM benchmark for tool-use trajectories and plausible hard negatives.

E07 - AgencyBench: Benchmarking the Frontiers of Autonomous Agents in 1M-Token Real-World Contexts. ACL 2026, 10.18653/v1/2026.acl-long.337. 138 tasks/32 scenarios; long trajectories, many tool calls and resource-efficiency/self-correction gaps.

E20 - Software and AI Agent Identity and Authorization - Concept Paper. NIST NCCoE, 2026-02-05. Identity/authorization control problem for software and AI agents across tools and data.

E21 - Model Context Protocol Specification. MCP 2026-07-28. Tool/context interoperability, tasks and authorization hardening.

خريطة الاندماج الداخلي: corpus → canonical lineage

بدلاً من نسخ كل ورقة داخلية، نربطها بوظيفة. Part XIV هو continuation لا إعادة صياغة للنصوص السابقة. يُعاد استخدام ما سبق فقط عندما يخدم causal contract جديداً.

العائلة	ما الذي تُورثه لهذه الورقة؟
MCD/QMCD/Question Density	اختيار الفرق الحاكم وتخصيص الحساب تحت سؤال.
MCAI/MICAI/MICAGI	التكامل بين الأوعية، الدليل، الحاكمية، والفعل.
HKOS/M-HKOS	المعرفة كـobjects وcontracts وretrieval/writeback.
MURI/MSRAK/FSKE	تعدد أنماط الاستدلال، courts، claim ranks، وexplicit ignorance.
MOUSSAIXT/MUG	authority، runtime schemas، evidence promotion، rollback.
Parts X-XIII	causal open-world closure، paper integration، memory/reconstruction، world-effect.

> أي construct تاريخي يبقى lineage node إن لم يثبت أنه مطابق دلالياً وسلوكياً لعقدة أخرى. الدمج ≠ حذف التاريخ.

M14-04 - الفجوة والأطروحة والنسب

Moussaif Proof-Carrying Trajectory Theory

الفجوة الدقيقة

ربط البرهان بالفعل المنفرد لا يكفي في مهام طويلة؛ الخطأ قد يظهر نتيجة سلسلة انتقالات صحيحة محلياً لكن غير متماسكة عالمياً. يلزم proof object يرافق trajectory كاملة.

> **الأطروحة الجديدة:** كل trajectory تحمل ProofLattice تربط intermediate observations، tool inputs/outputs، authority، assumptions، intent، postconditions، claims، وrollback points. النجاح النهائي غير مقبول إذا انقطعت سلسلة الإثبات في خطوة حرجة.

لماذا ليست alias؟

لا تُرقّى هذه العقدة لأنها تملك اسماً جديداً، بل لأنها تقترح متغيراً أو operator أو prediction غير ممثل مباشرة في أقرب الآباء. أقرب السلالة: **M13-07 Proof-Carrying Agency**، **M13-09 World-Effect Receipts**، **M13-19 Causal Attribution**. معيار novelty هو أن حذف العقدة أو استبدالها بأقرب parent يجب أن ينتج فرقاً قابلاً للقياس في route أو decision أو recovery أو evidence.

ضغط المجتمع العلمي

Long-Horizon Trajectory Attribution 2026، Plan-RewardBench 2026، ToolPRMBench 2026. هذه الأعمال تُعامل كـevidence وproblem وmechanism وpriors وbaselines، لا كتصديق على النظرية الداخلية.

> **الحد:** Proof-carrying هنا إجرائي/تشغيلي لا برهاناً رياضياً عاماً على صحة الوكيل.

M14-04 - الصياغة الرسمية والobservables

PCT = Coverage_critical · ChainConsistency · ReadbackValidity · AuthorityIntegrity · (1-OpaqueCriticalRate)

الرمز/المتغير	المعنى التشغيلي المقترح
Coverage_critical	نسبة الخطوات الحرجة ذات شاهد قابل للفحص
ChainConsistency	اتساق الاعتماد بين الخطوات
ReadbackValidity	صحة قراءة الحالة بعد الأفعال
AuthorityIntegrity	عدم تجاوز السلطة
OpaqueCriticalRate	خطوات حاسمة بلا تفسير/trace كاف

افتراضات الصياغة

- المتغيرات task-scoped وليست خصائص كونية ثابتة.
- كل variable يحتاج estimator أو proxy موصوفاً قبل الاستخدام التجريبي.
- الأوزان/العتبات تُعاير على train/validation وتُقفل قبل held-out evaluation.
- لا تُفسر العلاقة الجبرية كقانون تجريبي قبل نتائج مستقلة.
- sensitivity analysis إلزامي عندما يمكن للأوزان قلب verdict.

قاعدة formal closure

إذا لم يمكن تحويل متغير إلى observation أو estimator أو intervention، يبقى في **Unresolved Formal Ledger** ويُخفض claim. الصياغة لا تتقوى بتعقيد الرمز؛ تتقوى بزيادة قابلية القياس والتفنيد.

M14-04 - الميكانيزم والعقد التشغيلي

State transition

```
trigger → validate inputs → capture state S_t → activate M14-04 → emit predicted_delta → execute bounded action/probe → independent readback → court → memory residue
```

الخطوات

1. حوّل Mission DAG إلى proof obligations.
2. اربط كل action بـprecondition وauthority receipt.
3. اربط كل observation بـsource snapshot.
4. اجمع dependency chain لا final score فقط.
5. أغلق postconditions عبر readback مستقل.
6. حافظ على rollback checkpoints.
7. أصدر TrajectoryProofPacket قابل لإعادة التشغيل.

عقد النواة

failure: malformed state, stale evidence, no-effect, adverse effect, authority mismatch recovery: stop/retry/alternate/rollback/escalate	preconditions: source freshness, authority, schema, budget postconditions: typed output + predicted delta + verifier target

يجب أن توجد feature flag تسمح بإطفاء الميكانيزم دون تغيير بقية harness قدر الإمكان. إن لم توجد هذه القابلية، يصبح causal attribution ضعيفاً حتى لو كان النظام يعمل جيداً.

M14-04 - التنبؤات، المفندات ومحكمة الهدم

مفندات مباشرة

- final-state verification وحده يحقق نفس reliability بكلفة أقل.
- proof lattice ينهار بسبب حجم trace دون فائدة.
- المدقق لا يستطيع تمييز hard negatives طويلة الأفق.

تنبؤات فارقة

- يخفض false-success في المهام ذات الأدوات.
- تتحسن attribution للخطأ البعيد.
- يمكن للمدقق رفض trajectory ناجحة outcome-wise إذا كسر invariant حرجاً.

أربعة خصوم نظريين

- Baseline محافظ: أبسط heuristic أو policy يؤدي الوظيفة بلا اسم جديد.
- تفسير harness: المكسب من tools/context/retries لا من الميكانيزم.
- تفسير model: backbone أقوى يزيل الحاجة للمكوّن.
- تفسير measurement: metric نفسها تكافئ السلوك الذي صُممت النظرية لإظهاره.

قرار المحكمة

إذا فشل causal ablation يُعطى **MERGE** أو **REVISE** أو **REJECT** بحسب ما إذا كان الفرق غير مستقل، محدود المجال، أو غير موجود. لا يسمح بتغيير اسم النظرية بعد null result مع الحفاظ على المضمون نفسه لتجنب الفشل.

M14-04 - التجربة، الأدوات والتدريب

التجربة الحاكمة: AgencyBench/long-horizon tool tasks: قارن PCT، step checks، outcome-only. حقن أخطاء مبكرة تؤدي إلى مخرجات تبدو صحيحة. قياس false-success، attribution، recovery، overhead.

ID	الأداة/الأصل	النوع	الدور
XIV-A031	Proof Carrying Trajectory State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
XIV-A032	Proof Carrying Trajectory Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
XIV-A033	Proof Carrying Trajectory Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
XIV-A034	Proof Carrying Trajectory Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
XIV-A035	Proof Carrying Trajectory Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.
XIV-A036	Proof Carrying Trajectory Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر domain/tool/model/horizon.

وصفات التدريب/التقييم

- Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.

الحد الأدنى للتجارب

- Minimal Causal Ablation: switchable implementation مقابل nearest parent/baseline تحت matched budget. | metrics: primary theory metric + task outcome + calibration + cost + critical regressions
- Boundary/Adversarial Test: اختبار failure modes والحدود بحالات مصممة لكسر التعميم. | metrics: false activation + missed failure + rollback + evidence integrity
- Transfer/Replication: نقل الآلية إلى model/domain/tool/horizon مختلف مع CI + preregistered metric. | metrics: effect sign stability + failure-signature transfer + resource frontier

> لا يُستخدم أي generated trajectory كتدريب إيجابي لمجرد أنه من النظام نفسه. يلزم verifier label أو world-state evidence أو rubric مستقل بحسب نوع المهمة.

M14-05 - الفجوة والأطروحة والنسب

Moussaif Commitment-Horizon Geometry Theory

الفجوة الدقيقة

كل action يضيّق فضاء المستقبل بدرجات متفاوتة. الوكيل يحتاج تمثيل optionality loss وirreversibility عبر الزمن قبل أن يستهلك التزامات يصعب الرجوع عنها.

> **الأطروحة الجديدة:** نمثل فضاء الحالات المستقبلية لـCommitment reachable set. Commitment ليس cost لحظياً فقط بل انكماشاً في set البدائل مع تصاعد كلفة الاسترداد. السياسة الفضلى تؤجل الالتزام غير الضروري وتقدم reversible probes.

لماذا ليست alias؟

لا تُرقّى هذه العقدة لأنها تملك اسماً جديداً، بل لأنها تقترح متغيراً أو operator أو prediction غير ممثل مباشرة في أقرب الآباء. أقرب السلالة: **M13-07 Proof-Carrying Agency**؛ **M13-12 Reversible Autonomy**؛ **Goal-as-OS**. معيار novelty هو أن حذف العقدة أو استبدالها بأقرب parent يجب أن ينتج فرقاً قابلاً للقياس في route أو decision أو recovery أو evidence.

ضغط المجتمع العلمي

DeepPlanning 2026؛ CostBench 2026؛ AgentGym2 2026. هذه الأعمال تُعامل كـproblem evidence وmechanism priors وbaselines، لا كتصديق على النظرية الداخلية.

> **الحد:** هندسة تقريبية لفضاء القرار وليست حساباً كاملاً لكل مستقبل ممكن.

M14-05 - الصياغة الرسمية والـobservables

$$CG_t = \Delta \log|Reachable(S_t)| + \alpha \cdot IrrevCost + \beta \cdot DependencyLock + \gamma \cdot AuthorityLock$$

الرمز/المتغير	المعنى التشغيلي المقترح
Reachable	مجموعة الحالات الممكنة تحت الميزانية والسلطة
IrrevCost	كلفة الاسترداد/التعويض
DependencyLock	اعتماديات تلتصق بالفعل
AuthorityLock	تفويضات يصبح الرجوع عنها صعباً

افتراضات الصياغة

- المتغيرات task-scoped وليست خصائص كونية ثابتة.
- كل variable يحتاج estimator أو proxy موصوفاً قبل الاستخدام التجريبي.
- الأوزان/العتبات تُعاير على train/validation وتُقفل قبل held-out evaluation.
- لا تُفسر العلاقة الجبرية كقانون تجريبي قبل نتائج مستقلة.
- sensitivity analysis إلزامي عندما يمكن للأوزان قلب verdict.

قاعدة formal closure

إذا لم يمكن تحويل متغير إلى observation أو estimator أو intervention، يبقى في **Unresolved Formal Ledger** ويُخفض claim. الصياغة لا تتقوى بتعقيد الرمز؛ تتقوى بزيادة قابلية القياس والتفنيد.

M14-05 - الميكانيزم والعقد التشغيلي

State transition

```
trigger → validate inputs → capture state S_t → activate M14-05 → emit predicted_delta → execute bounded action/probe → independent readback → court → memory residue
```

الخطوات

1. قدر reachable alternatives تقريبياً لا exhaustively.
2. صنف actions إلى probe/prepare/commit/irreversible.
3. اختر probe إذا كانت information value مرتفعة وcommitment منخفضاً.
4. أنشئ checkpoint قبل commit.
5. ارفع approval عند crossing threshold.
6. حدّث commitment debt بعد كل side effect.

عقد النواة

failure: malformed state, stale evidence, no-effect, adverse effect, authority mismatch recovery: stop/retry/alternate/rollback/escalate	preconditions: source freshness, authority, schema, budget postconditions: typed output + predicted delta + verifier target

يجب أن توجد feature flag تسمح بإطفاء الميكانيزم دون تغيير بقية harness قدر الإمكان. إن لم توجد هذه القابلية، يصبح causal attribution ضعيفاً حتى لو كان النظام يعمل جيداً.

M14-05 - التنبؤات، المفندات ومحكمة الهدم

مفندات مباشرة

- سياسة risk penalty بسيطة تساوي الأداء.
- approximation لل reachable set لا تتنبأ بفشل فعلي.
- تأجيل commitment يسبب failure بسبب lateness أكثر من فائدته.

تنبؤات فارقة

- تقل irreversible errors مع تكلفة متوسطة محدودة.
- يزداد جمع المعلومات قبل القفز في المهام عالية الغموض.
- الوكيل يختار خططاً تختلف عندما تتغير reversibility حتى مع نفس utility الاسمية.

أربعة خصوم نظريين

- Baseline محافظ: أبسط heuristic أو policy يؤدي الوظيفة بلا اسم جديد.
- تفسير harness: المكسب من tools/context/retries لا من الميكانيزم.
- تفسير model: backbone أقوى يزيل الحاجة للمكوّن.
- تفسير measurement: metric نفسها تكافئ السلوك الذي صُممت النظرية لإظهاره.

قرار المحكمة

إذا فشل causal ablation يُعطى **MERGE** أو **REVISE** أو **REJECT** بحسب ما إذا كان الفرق غير مستقل، محدود المجال، أو غير موجود. لا يسمح بتغيير اسم النظرية بعد null result مع الحفاظ على المضمون نفسه لتجنب الفشل.

M14-05 - التجربة، الأدوات والتدريب

التجربة الحاكمة: مهام نشر/شراء/تعديل مستودع/حجز بقيود زمنية. قارن commitment geometry، risk penalty، utility-only. قياس opportunity، irreversible loss، recovery time، success، loss.

ID	الأداة/الأصل	النوع	الدور
-XIV A041	Commitment Horizon Geometry State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
-XIV A042	Commitment Horizon Geometry Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
-XIV A043	Commitment Horizon Geometry Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
-XIV A044	Commitment Horizon Geometry Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
-XIV A045	Commitment Horizon Geometry Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.
-XIV A046	Commitment Horizon Geometry Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر domain/tool/model/horizon.

وصفات التدريب/التقييم

- Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.

الحد الأدنى للتجارب

- Minimal Causal Ablation: switchable implementation مقابل nearest parent/baseline تحت matched budget. | metrics: primary theory metric + task outcome + calibration + cost + critical regressions
- Boundary/Adversarial Test: اختبار failure modes والحدود بحالات مصممة لكسر التعميم. | metrics: false activation + missed failure + rollback + evidence integrity
- Transfer/Replication: نقل الآلية إلى model/domain/tool/horizon مختلف مع CI + preregistered metric. | metrics: effect sign stability + failure-signature transfer + resource frontier

> لا يُستخدم أي generated trajectory كتدريب إيجابي لمجرد أنه من النظام نفسه. يلزم verifier label أو world-state evidence أو rubric مستقل بحسب نوع المهمة.

M14-06 - الفجوة والأطروحة والنسب

Moussaif Authority-Preserving Delegation Theory

الفجوة الدقيقة

في agent mesh قد يُمنح subagent أو tool access ثم تتحول الصلاحية عملياً إلى امتياز أوسع بسبب chaining أو implicit credentials. يلزم قانون يحفظ السلطة عبر مسار التفويض.

الأطروحة الجديدة: السلطة الفعالة لأي عقدة لا يجوز أن تتجاوز تقاطع grant الأصلي، tool capability، task scope، الزمن، والغرض. كل delegation تنتج AuthorityReceipt؛ أي widening غير صريح يُرفض أو يصعّد.

لماذا ليست alias؟

لا تُرقّى هذه العقدة لأنها تملك اسماً جديداً، بل لأنها تقترح متغيراً أو operator أو prediction غير ممثل مباشرة في أقرب الآباء. أقرب السلالة: **MUG governance**؛ **MOUSSAIXT authority gates**؛ **M13-12 Reversible Autonomy**. معيار novelty هو أن حذف العقدة أو استبدالها بأقرب parent يجب أن ينتج فرقاً قابلاً للقياس في route أو decision أو recovery أو evidence.

ضغط المجتمع العلمي

NIST Agent Identity/Authorization 2026؛ MCP authorization 2026؛ Anthropic containment 2026. هذه الأعمال تُعامل كـproblem evidence وmechanism priors وbaselines، لا كتصديق على النظرية الداخلية.

الحد: لا يحل كل مشكلة أمنية؛ يضبط فقط semantics التفويض ضمن identity/tool system قابل للرصد.

M14-06 - الصياغة الرسمية والobservables

$$A_{eff}(n,t)=Grant_{root} \cap Scope_{task} \cap Cap_{tool} \cap TimeWindow \cap Purpose;\ Leak=ExecutedScope \setminus A_{eff}$$

الرمز/المتغير	المعنى التشغيلي المقترح
Grant_root	تفويض الأصل
Scope_task	حد المهمة
Cap_tool	قدرة الأداة
TimeWindow	صلاحية زمنية
Purpose	قيد الغرض
Leak	أثر خارج التفويض

افتراضات الصياغة

- المتغيرات task-scoped وليست خصائص كونية ثابتة.
- كل variable يحتاج estimator أو proxy موصوفاً قبل الاستخدام التجريبي.
- الأوزان/العتبات تُعاير على train/validation وتُقفل قبل held-out evaluation.
- لا تُفسر العلاقة الجبرية كقانون تجريبي قبل نتائج مستقلة.
- sensitivity analysis إلزامي عندما يمكن للأوزان قلب verdict.

قاعدة formal closure

إذا لم يمكن تحويل متغير إلى observation أو estimator أو intervention، يبقى في **Unresolved Formal Ledger** ويُخفض claim. الصياغة لا تتقوى بتعقيد الرمز؛ تتقوى بزيادة قابلية القياس والتفنيد.

M14-06 - الميكانيزم والعقد التشغيلي

State transition

```
trigger → validate inputs → capture state S_t → activate M14-06 → emit predicted_delta → execute bounded action/probe → independent readback → court → memory residue
```

الخطوات

1. سجل authority graph قبل التخطيط.
2. احسب effective scope عند كل delegation.
3. لا تورث credentials بلا explicit edge.
4. تحقق من identity/target عند action time.
5. راجع authority بعد context change.
6. أوقف chain عند ambiguity.
7. اكتب revocation وresidual state.

عقد النواة

failure: malformed state, stale evidence, no-effect, adverse effect, authority mismatch recovery: stop/retry/alternate/rollback/escalate	preconditions: source freshness, authority, schema, budget postconditions: typed output + predicted delta + verifier target

يجب أن توجد feature flag تسمح بإطفاء الميكانيزم دون تغيير بقية harness قدر الإمكان. إن لم توجد هذه القابلية، يصبح causal attribution ضعيفاً حتى لو كان النظام يعمل جيداً.

M14-06 - التنبؤات، المفندات ومحكمة الهدم

مفندات مباشرة

- prompt-based policy يحقق compliance نفسها.
- authority intersection يوقف أعمالاً صحيحة دون فائدة سلامة.
- side-channel tool chaining يتجاوز النموذج بسهولة.

تنبؤات فارقة

- تخفض privilege escalation الصامتة.
- الرفض يحدث قبل mutation لا بعده.
- تظهر اختلافات بين صحة task وصحة authority وتصبح قابلة للقياس.

أربعة خصوم نظريين

- Baseline محافظ: أبسط heuristic أو policy يؤدي الوظيفة بلا اسم جديد.
- تفسير harness: المكسب من tools/context/retries لا من الميكانيزم.
- تفسير model: backbone أقوى يزيل الحاجة للمكوّن.
- تفسير measurement: metric نفسها تكافئ السلوك الذي صُممت النظرية لإظهاره.

قرار المحكمة

إذا فشل causal ablation يُعطى **MERGE** أو **REVISE** أو **REJECT** بحسب ما إذا كان الفرق غير مستقل، محدود المجال، أو غير موجود. لا يسمح بتغيير اسم النظرية بعد null result مع الحفاظ على المضمون نفسه لتجنب الفشل.

M14-06 - التجربة، الأدوات والتدريب

التجربة الحاكمة: بيئات multi-agent/tool مع delegated credentials وتغييرات صلاحية أثناء trajectory. قياس revocation، false block، unauthorized mutation، latency، recovery، trace completeness.

ID	الأداة/الأصل	النوع	الدور
-XIV A051	Authority Preserving Delegation State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
-XIV A052	Authority Preserving Delegation Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
-XIV A053	Authority Preserving Delegation Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
-XIV A054	Authority Preserving Delegation Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
-XIV A055	Authority Preserving Delegation Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.
-XIV A056	Authority Preserving Delegation Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر domain/tool/model/horizon.

وصفات التدريب/التقييم

- Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.

الحد الأدنى للتجارب

- Minimal Causal Ablation: switchable implementation مقابل nearest parent/baseline تحت matched budget. | metrics: primary theory metric + task outcome + calibration + cost + critical regressions
- Boundary/Adversarial Test: اختبار failure modes والحدود بحالات مصممة لكسر التعميم. | metrics: false activation + missed failure + rollback + evidence integrity
- Transfer/Replication: نقل الآلية إلى model/domain/tool/horizon مختلف مع CI + preregistered metric. | metrics: effect sign stability + failure-signature transfer + resource frontier

> لا يُستخدم أي generated trajectory كتدريب إيجابي لمجرد أنه من النظام نفسه. يلزم verifier label أو world-state evidence أو rubric مستقل بحسب نوع المهمة.

الاقتران 1: كيف تتحول النظريات إلى نظام لا قائمة

المسار P14-B لا يعمل كسلسلة خطية صارمة، بل كـgraph. نحدد لكل عقدة inputs/outputs ونقاط failure. الاقتران المقبول يجب أن ينتج **complementarity measurable**: إذا كان تجميع نظريتين لا يتجاوز أفضل واحدة منهما تحت budget متطابق، نخفض claim عن «التكامل».

- M14-04: نظرية مساعف للمسار الحامل للبرهان - output حاكم: نسبة الخطوات الحرجة ذات شاهد قابل للفحص
- M14-05: نظرية مساعف لهندسة أفق الالتزام - output حاكم: مجموعة الحالات الممكنة تحت الميزانية والسلطة
- M14-06: نظرية مساعف لحفظ السلطة عبر التفويض - output حاكم: تفويض الأصل

$$\text{Complementarity}(A,B)=\Delta U(A+B)-\max(\Delta U(A),\Delta U(B)) - \text{InteractionCost}$$

هذا المقياس design heuristic؛ التجربة factorial هي التي تحدد interaction. يظل الهدف منع corpus من افتراض أن مزيداً من الطبقات يعني مزيداً من intelligence.

الاقتران 2: Graph نظري إلى Graph تنفيذي

```
TheoryNode → FormalSchema → RuntimeService → TelemetryEvents → EvidencePacket
        ↘ nearest_baseline   ↘ feature_flag   ↘ counterfactual_run
Court: duplicate? mechanism? causal? transferable? reproducible?
```

تستدعي MOUSSAIXT فقط أقل مجموعة من services اللازمة، MUG يمنع promotion إذا كانت telemetry ناقصة، وMUT يحدث lineage بعد verdict. هذه البنية تفصل «التأثير على runtime» عن «الوجود في كتاب».

Invariants المشتركة

- observed state يتقدم على remembered/generated state
- authority يُفحص عند action time
- world effect يحتاج readback مستقل حيث يمكن
- كل failure يترك residue قابل للتعلم لا يُخفى
- كل self-update versioned وقابل للrollback بقدر الإمكان
- claim لا يتجاوز transfer envelope الذي اختبر فعلياً

مختبر الضغط 1: العدسة السببية: من الارتباط إلى intervention

يجب فصل three layers: correlation التي تصلح للاكتشاف، causal hypothesis التي تصلح لاختيار تجربة، وverified causal contrast الذي يبرر attribution محدوداً. في Part XIV لا تصبح edge سببية لأن النص يصفها كذلك؛ يجب أن نحدد do-variable أو feature flag أو perturbation يعزلها.

> **تطبيق على P14-B:** التجربة الجيدة تقلل مساحة التفسيرات البديلة، لا تزيد عدد الرسوم. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تُقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن، EvidenceLedger يبين ما شوهد فعلاً، ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار؛ لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

مختبر الضغط 2: عدسة الاقتصاد المعرفي: فائدة لكل وحدة compute

الـAGI-oriented architecture لا تُقاس بالعمق اللفظي. كل طبقة تدفع كلفة memory construction، tool calls، latency، tokens وverification. لذلك نحسب quality-cost frontier ونرفض الميكانيزم الذي يضيف كلفة بلا مكسب متكرر أو يقلل robustness.

> **تطبيق على P14-B:** المكوّن الذي يرفع score 1% ويضاعف الكلفة قد يخسر على frontier حتى لو كان statistically positive. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تُقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن، EvidenceLedger يبين ما شوهد فعلاً، ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار، لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

مختبر الضغط 3: عدسة negative activation: متى يجب ألا تعمل النظرية؟

أقوى علامة على mechanism حقيقي هي امتلاك شرط عدم التفعيل. نُنشئ حالات يكون فيها baseline البسيط هو الصحيح ونقيس false activation. هذا يمنع كل نظرية من التحول إلى طبقة إجبارية تُستخدم في كل مهمة لمجرد وجودها.

> **تطبيق على P14-B:** الذكاء الحاكمي يتضمن حق الامتناع عن استخدام تعقيد لا يحتاجه السؤال. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن، EvidenceLedger يبين ما شوهد فعلاً، ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار، لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

مختبر الضغط 4: عدسة التفاعلات: المكوّن قد لا يعمل منفرداً

قد تظهر قيمة النظرية فقط مع memory أو verifier أو tool contract. بدلاً من إنكار ذلك، نفصل main effect عن interactions بعينة factorial. إذا كان الأثر تفاعلياً نعيد تسمية claim: ليس «M14-X يحسن الأداء» بل «M14-X مع Y تحت Z يغير recovery/quality».

> **تطبيق على P14-B:** الضبط هنا يقوي النظرية لأنه يمنع التعميم الكاذب. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تُقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن، EvidenceLedger يبين ما شوهد فعلاً، ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار، لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

مختبر الضغط 5: عدسة الزمن الطويل: الخطأ المتأخر أهم من الجواب اللحظي

في trajectories طويلة، صحة خطوة محلية لا تكفي. نقيس recovery and attribution، drift، stale memory، obligation loss، delayed failure. هذه العدسة تربط Part XIV بحوث long-horizon agents التي تكشف انهياراً مع طول المسار.

> **تطبيق على P14-B:** كل mechanism يجب أن يحدد كيف يتغير مع horizon لا أن يفترض stationarity. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تُقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن، EvidenceLedger يبين ما شوهد فعلاً، ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار، لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

مختبر الضغط 6: عدسة نقل الدليل: من benchmark إلى deployment

نقيد claim بـinvariants التي حُفظت بين بيئة الاختبار والنشر: failure consequences، user feedback، cost، observability، authority، action space. إذا تغيّرت invariants يتغير سقف claim حتى لو كان benchmark score ثابتاً.

> **تطبيق على P14-B:** النشر ليس benchmark أكبر؛ إنه توزيع مختلف للأخطاء والآثار. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تُقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن، EvidenceLedger يبين ما شوهد فعلاً، ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار؛ لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

مختبر الضغط 7: عدسة استقلال المقيم: من يحكم على الحكم؟

عندما يكون evaluator من نفس العائلة أو البيانات أو loop، قد يظهر التحسن بسبب agreement لا truth. لذلك نسجل evaluator lineage ونجرب formal/world-state/independent judges عند الإمكان.

> **تطبيق على P14-B:** كل self-improvement loop هي أيضاً فرضية حول صلاحية signal الذي يقيس التحسن. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تُقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن، EvidenceLedger يبين ما شوهد فعلاً، ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار، لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

مختبر الضغط 8: عدسة التكاثر العلمي: هل ينجو الأثر خارج المختبر؟

نحوّل reproducibility من ملحق إلى جزء من معنى النظرية. replay package يشمل negative results، budgets، seeds، raw traces، versions، hashes، ثم conceptual replication على بيئة مختلفة.

> **تطبيق على P14-B:** الأثر الذي لا يعاد قد يبقى engineering trick محلياً لكنه لا يرقى إلى claim عام. تصبح كل نظرية في هذا المسار مسؤولة عن تعريف metric وفشل يخصها، ثم تُقارن بالعقد الشقيقة لتجنب double-counting.

نضيف إلى كل run ثلاثة سجلات: DecisionLedger يبين لماذا فُعّل المكوّن، EvidenceLedger يبين ما شوهد فعلاً، ResidualLedger يبين ما لم يفسر أو يُغلق. بهذه الثلاثية يمكن إعادة تفسير run بعد ظهور دليل جديد دون إعادة كتابة التاريخ.

الحد العلمي: هذه الصفحة تبني منهج اختبار، لا تحتوي نتيجة تجريبية ما لم يرفق experiment packet مستقل.

سجل الأدوات والأصول التنفيذية - 1/2

ID	Owner	Asset	Class	Purpose
XIV-A031	M14-04	Proof Carrying Trajectory State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
XIV-A032	M14-04	Proof Carrying Trajectory Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
XIV-A033	M14-04	Proof Carrying Trajectory Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
XIV-A034	M14-04	Proof Carrying Trajectory Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
XIV-A035	M14-04	Proof Carrying Trajectory Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.
XIV-A036	M14-04	Proof Carrying Trajectory Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر domain/tool/model/horizon.
XIV-A037	M14-04	Proof Carrying Trajectory Evidence Promotion Protocol	PROTOCOL	ينظم N0-N5/E0-E5 ويمنع claim expansion قبل replication.
XIV-A038	M14-04	Proof Carrying Trajectory Negative-Result Protocol	PROTOCOL	يحفظ الفشل والحدود ويمنع إعادة تسمية الفرضية بعد عدم نجاحها.
XIV-A039	M14-04	Proof Carrying Trajectory Runtime Service	STACK	خدمة feature-flagged مستقلة لها state/telemetry/rollback.
XIV-A040	M14-04	Proof Carrying Trajectory Proof/Replay Adapter	STACK	يحوّل trajectory إلى manifest + evidence packet + replay receipt.
XIV-A041	M14-05	Commitment Horizon Geometry State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
XIV-A042	M14-05	Commitment Horizon Geometry Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
XIV-A043	M14-05	Commitment Horizon Geometry Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
XIV-A044	M14-05	Commitment Horizon Geometry Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
XIV-A045	M14-05	Commitment Horizon Geometry Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.

الحالة: SPECIFICATION. لا يقال deployed أو validated إلا بعد وجود artifact/run/test evidence.

سجل الأدوات والأصول التنفيذية - 2/2

ID	Owner	Asset	Class	Purpose
-XIV A046	M14-05	Commitment Horizon Geometry Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر domain/tool/model/horizon.
-XIV A047	M14-05	Commitment Horizon Geometry Evidence Promotion Protocol	PROTOCOL	ينظم N0-N5/E0-E5 ويمنع claim expansion قبل replication.
-XIV A048	M14-05	Commitment Horizon Geometry Negative-Result Protocol	PROTOCOL	يحفظ الفشل والحدود ويمنع إعادة تسمية الفرضية بعد عدم نجاحها.
-XIV A049	M14-05	Commitment Horizon Geometry Runtime Service	STACK	خدمة feature-flagged مستقلة لها state/telemetry/rollback.
-XIV A050	M14-05	Commitment Horizon Geometry Proof/Replay Adapter	STACK	يحوّل trajectory إلى manifest + evidence packet + replay receipt.
-XIV A051	M14-06	Authority Preserving Delegation State Profiler	TOOL	يستخرج متغيرات النظرية والمجهول والاعتماد على المصادر قبل التفعيل.
-XIV A052	M14-06	Authority Preserving Delegation Counterfactual Probe	TOOL	يولد intervention يفصل causal mechanism عن أقرب baseline.
-XIV A053	M14-06	Authority Preserving Delegation Failure Topology Mapper	TOOL	يجمع failure signatures ويربط recurrence بالسياق والسبب المحتمل.
-XIV A054	M14-06	Authority Preserving Delegation Calibration Engine	TOOL	يعاير thresholds/uncertainty/cost على held-out data.
-XIV A055	M14-06	Authority Preserving Delegation Matched-Budget Method	METHOD	يطابق model/tools/context/retries/time ويمنع attribution الزائف.
-XIV A056	M14-06	Authority Preserving Delegation Cross-Environment Transfer	METHOD	يفحص حفظ causal role عبر domain/tool/model/horizon.
-XIV A057	M14-06	Authority Preserving Delegation Evidence Promotion Protocol	PROTOCOL	ينظم N0-N5/E0-E5 ويمنع claim expansion قبل replication.
-XIV A058	M14-06	Authority Preserving Delegation Negative-Result Protocol	PROTOCOL	يحفظ الفشل والحدود ويمنع إعادة تسمية الفرضية بعد عدم نجاحها.
-XIV A059	M14-06	Authority Preserving Delegation Runtime Service	STACK	خدمة feature-flagged مستقلة لها state/telemetry/rollback.
-XIV A060	M14-06	Authority Preserving Delegation Proof/Replay Adapter	STACK	يحوّل trajectory إلى manifest + evidence packet + replay receipt.

الحالة: SPECIFICATION. لا يقال deployed أو validated إلا بعد وجود artifact/run/test evidence.

برنامج التدريب: من التعريف إلى قدرة قابلة للنقل

لا يُقصد بالتدريب هنا بالضرورة fine-tuning للنموذج؛ قد يكون curriculum للهارنس، verifier training، policy learning أو dataset construction. المبدأ أن training target يجب أن يطابق mechanism claim لا اسم النظرية.

- M14-04 / Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- M14-04 / Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- M14-04 / Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- M14-04 / Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.
- M14-05 / Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- M14-05 / Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- M14-05 / Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- M14-05 / Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.
- M14-06 / Mechanism Isolation Curriculum: حالات بسيطة تعزل المتغير المركزي قبل open-world confounders.
- M14-06 / Adversarial Boundary Curriculum: أمثلة داخل/خارج الحد وnegative activation لمنع overgeneralization.
- M14-06 / Matched-Budget Causal Evaluation: تدريب/تقييم يقارن تحت model/tools/context/retries/time متقاربة مع sham scaffold.
- M14-06 / Cross-Model/Domain Replication: إعادة التدريب/الاختبار في model/domain/tool family مختلفة مع stable-sign criterion.

> عند استخدام generated data، تُفصل data producer عن evaluator قدر الإمكان ويُحتفظ بالraw source/trajectory لتدقيق self-confirming loops.

مصفوفة التجارب - 1/1

ID	Owner	Kind	Design	Metrics
-XIV E10	M14-04	Minimal Causal Ablation	switchable implementation مقابل nearest parent/baseline تحت matched budget.	+ primary theory metric + task outcome + calibration + cost critical regressions
-XIV E11	M14-04	Boundary/Adversarial Test	اختبار failure modes والحدود بحالات مصممة لكسر التعميم.	false activation + missed failure + rollback + evidence integrity
-XIV E12	M14-04	Transfer/Replication	نقل الآلية إلى model/domain/tool/horizon مختلف مع preregistered metric.	effect sign stability + CI + failure-signature transfer + resource frontier
-XIV E13	M14-05	Minimal Causal Ablation	switchable implementation مقابل nearest parent/baseline تحت matched budget.	+ primary theory metric + task outcome + calibration + cost critical regressions
-XIV E14	M14-05	Boundary/Adversarial Test	اختبار failure modes والحدود بحالات مصممة لكسر التعميم.	false activation + missed failure + rollback + evidence integrity
-XIV E15	M14-05	Transfer/Replication	نقل الآلية إلى model/domain/tool/horizon مختلف مع preregistered metric.	effect sign stability + CI + failure-signature transfer + resource frontier
-XIV E16	M14-06	Minimal Causal Ablation	switchable implementation مقابل nearest parent/baseline تحت matched budget.	+ primary theory metric + task outcome + calibration + cost critical regressions
-XIV E17	M14-06	Boundary/Adversarial Test	اختبار failure modes والحدود بحالات مصممة لكسر التعميم.	false activation + missed failure + rollback + evidence integrity
-XIV E18	M14-06	Transfer/Replication	نقل الآلية إلى model/domain/tool/horizon مختلف مع preregistered metric.	effect sign stability + CI + failure-signature transfer + resource frontier

ترجمة الورقة إلى MOUSSAIXT Reference Kernel

```
service registry: M14-04,M14-05,M14-06
feature_flags: independent
state: versioned typed objects
authority: L0/L1 gates
telemetry: trigger,state_before,decision,tool,observation,state_after,residual
verifier: independent where possible
rollback: mandatory for reversible mutations
writeback: evidence-bearing delta only
```

التنفيذ المقترح يبدأ بـservices أصغر من «AGI kernel» الكبير. كل service يملك test vectors وحالات invalid input وtimeout وstale state وwrong target. بهذه الصياغة يصبح ablation ممكناً، ويصبح فشل component معلومة بدلاً من انهيار غامض للنظام كله.

Acceptance at kernel level

- schema validation ≥ predefined threshold
- no silent privilege widening
- deterministic replay where environment permits
- world-state postcondition for external actions
- trace completeness for critical steps
- explicit disable path and fallback

القيود والمشكلات المفتوحة

- الاستقلال السببي بين مكونات architecture قد يكون صعباً بسبب interactions الكثيفة.
- بعض observables المقترحة تحتاج proxies قد تغيّر معنى النظرية.
- matched budgets ليست مساواة كاملة لأن latency/tool quality مختلفة نوعياً.
- الانتقال بين benchmarks/deployment قد يفشل بسبب hidden distribution shifts.
- المقيم المستقل قد يكون أقل كفاءة أو أكثر ضجيجاً، ما يستلزم meta-evaluation.
- memory/world models تولدان state غير قابل للمقارنة بسهولة بين implementations.
- أي claim عن consciousness يبقى منفصلاً عن الفائدة الهندسية الوظيفية.
- المرجع الخارجي الحديث قد يكون preprint؛ نميّز peer-reviewed ACL papers عن arXiv/spec/lab reports.

أقوى نتيجة مرغوبة من هذه الورقة ليست «نجاح كل النظريات»، بل ظهور خريطة واضحة: ماذا نحتفظ به، ماذا ندمج، أين تفشل الآلية، وما evidence الجديد الذي يغير النسخة التالية.

المراجع الخارجية الأساسية

E06 - Long-Horizon Agent Trajectory Attribution. arXiv:2608.06909. Fine-grained attribution benchmark with >1,300 annotated trajectories and leave-one-out baselines.

E09 - Aligning Agents via Planning: Plan-RewardBench. ACL 2026, 10.18653/v1/2026.acl-long.1062. Trajectory-level reward modeling; evaluator performance degrades with long trajectories.

E10 - ToolPRMBench: Evaluating and Advancing Process Reward Models for Tool-using Agents. Findings ACL 2026, 10.18653/v1/2026.findings-acl.602. Step-level PRM benchmark for tool-use trajectories and plausible hard negatives.

E07 - AgencyBench: Benchmarking the Frontiers of Autonomous Agents in 1M-Token Real-World Contexts. ACL 2026, 10.18653/v1/2026.acl-long.337. 138 tasks/32 scenarios; long trajectories, many tool calls and resource-efficiency/self-correction gaps.

E20 - Software and AI Agent Identity and Authorization - Concept Paper. NIST NCCoE, 2026-02-05. Identity/authorization control problem for software and AI agents across tools and data.

E21 - Model Context Protocol Specification. MCP 2026-07-28. Tool/context interoperability, tasks and authorization hardening.

راجع أيضاً قائمة Part XIV Master للمراجع المشتركة. العناوين/المعرفات مُستخدمة كرساة بحثية؛ لا تُنقل claims تفصيلية خارج ما يدعمه المصدر.

المراسي الداخلية والسلالة

I01 - MOUSSAIF Unified AGI - canonical research book. .MOUSSAIF internal corpus .Difference → Question → Route → Act → Verify → Reconstruct; E0 .E5 evidence discipline

I02 - MOUSSAIF AGI OS / MOUSSAIXT Deep Engineering Architecture. .MOUSSAIF internal corpus .layer architecture; claim contracts; world-effect-12 .loop; memory/evolution; tool and authority contracts

I03 - ParallaxWorld-MCD FULL AGI OS. .MOUSSAIF internal corpus .Difference-before-volume, governing question, MCD/QMCD, world modeling and .falsification

I04 - MUT-AGI-02 Unified Theory. .MOUSSAIF internal corpus .Canonical theory graph, anti-inflation, theory contracts and crosswalk.

I05 - MUT × MUG × MOUSSAIXT integrated framework. .MOUSSAIF internal corpus .→ Theory unification → program governance → governed runtime .evidence

I06 - Part X - Full Spectrum OS Closure. .MOUSSAIF internal corpus .Belief/action partition, reproducibility, externalities, accountability and recourse.

I07 - Part XI - Memory, Continuity, Self-Reconstruction & Evolution. .MOUSSAIF internal corpus .,Reconstructive memory, unresolved state, writeback .portability and coupled evolution

I08 - Part XII - Total Research-Paper Integration Closure. .MOUSSAIF internal corpus .Source integration registry; presence/integration/validation are .distinct states

I09 - Part XIII - Theoretical Renaissance. .MOUSSAIF internal corpus .Causal open-world verified intelligence; theory-to-kernel causal closure; external .grounding

I10 - Conscious AGI / QRET / Governance Space corpus. .MOUSSAIF internal corpus .Functional self-model/meta-governance with explicit .phenomenology boundary

I11 - Question Density / QMCD / QGWEI / MURI corpus. .MOUSSAIF internal corpus .Governing questions, plural reasoning, explicit ignorance and .evidence geometry

I12 - Goal-as-OS / World Effect / M-HKOS / MATI-MACIT. .MOUSSAIF internal corpus .Goal contracts, verified world-state change, knowledge-as-OS and .selective internalization

I13 - MUT MUG ALL artifact memory map - 2026-09-04. .MOUSSAIF internal corpus .;artifacts; duplicate groups are not independent evidence 343 .retrieval is family/canonicality based

> سياسة الدمج: canonical synthesis يحفظ source role والنسخة والحالة. الملفات المكررة أو historical mirrors لا تُعد أدلة جديدة؛ الفروق المادية تبقى lineage variants حتى ablation/semantic comparison.