RECONSTRUCTED_FROM_INDEXED_TEXT - نسخة أرشيفية معاد بناؤها من النص المفهرس

<PARSED TEXT FOR PAGE: 1 / 25>
PAPER-XIV-02 · MOUSSAIF AGI RESEARCH · 2026
المسارات الحاملة للبرهان: هندسة الالتزام وحفظ السلطة في
الوكالة طويلة الأفق
Proof-Carrying Trajectories: Commitment Geometry and Authority Conservation in Long-Horizon Agency
سلامة الفعل لا تكفي إذا كانت trajectory كاملة تنتهك الهدف أو السلطة. نقترح state proof متراكماً للمسار، هندسة زمنية للالتزام، وقانون حفظ للسلطة أثناء التفويض عبر /tools/agents
protocols.
MUT lineage → formal construct → switchable mechanism → matched-budget test → world/evidence readback → replication
Research specification / theory paper. Candidate contributions; not empirical proof of AGI.
1
<PARSED TEXT FOR PAGE: 2 / 25>
ABSTRACT الملخص والادعاء المركزي
سلامة الفعل لا تكفي إذا كانت trajectory كاملة تنتهك الهدف أو السلطة. نقترح state proof متراكماً للمسار، هندسة زمنية للالتزام، وقانون حفظ للسلطة أثناء التفويض عبر protocols/tools/agents.
تبني الورقة على XII Part الذي فصل presence عن integration، وXIII Part الذي صاغ verified world-open causal closure، ثم تستخدم XIV Part لإغلاق طبقة محددة بآليات قابلة للتعطيل والاختبار. الأدبيات الخارجية تُستخدم كـart prior وتصميم pressure failure/benchmark، ولا تُستخدم كدليل أن MOUSSAIF حقق نتائجها.
Contributions
توسيع Proof-Carrying Agency إلى Proof-Carrying Trajectory ذات invariants وobligations متراكمة.
تعريف Commitment Horizon وCritical Rollback Time إلى الخيار تحول قبل lock-in.
صياغة Conservation Authority Delegated كـintersection لا كنسخ للصلاحيات.
ربط MCP/A2A/IAM بطبقة semantics مستقلة للauthority/state/evidence.
Scientific value = distinct causal hypothesis × falsifiability × reproducibility × transfer
1.
2.
3.
4.
2
<PARSED TEXT FOR PAGE: 3 / 25>
CONTRIBUTIONS المساهمات وحدود الأسبقية
Status Contribution
ذات invariants وobligations متراكمة. new synthesis within MOUSSAIF; requires literature novelty audit and experiments توسيع Proof-Carrying Agency إلى Proof-Carrying Trajectory
Critical إلى الخيار تحول قبل new synthesis within MOUSSAIF; requires literature novelty audit and experiments .lock-in تعريف Commitment Horizon وRollback Time
كintersection لا كنسخ للصلاحيات. new synthesis within MOUSSAIF; requires literature novelty audit and experiments صياغة Delegated Authority Conservation
للauthority/state/evidence مستقلة new synthesis within MOUSSAIF; requires literature novelty audit and experiments. ربط MCP/A2A/IAM بطبقة semantics
لا تدعي هذه الورقة أن المصطلحات أو كل المكونات بلا سوابق عالمية؛ المجتمع العلمي يملك خطوطاً قوية في /protocols/control. novelty المقترحة تقع في تركيبها إلى benchmarks/memory/models world/reasoning agentic
داخل MUT/MUG/MOUSSAIXT. المشتركة والاختبارات formal distinctions وفي closure law.
DISCIPLINE EVIDENCE منهج المصدر والدليل
reported evidence؛ يظل vendor report أو engineering inference → synthesis → internal source-preserved artifact
الأدلة ترتيب: source paper/system spec المباشر → hypothesis. system card
لا يصبح preprint arXiv حقيقة نهائية؛ paper internal لا يثبت success runtime. كل claim في هذه الورقة يجب أن يستطيع الرجوع إلى role source أو أن يُوسم بوضوح كاقتراح.
External

prior art، benchmark design، failure modes، protocol facts.
Internal
MUT lineage، MUG placement، MOUSSAIXT contracts، historical theories.
3
<PARSED TEXT FOR PAGE: 4 / 25>
1/4 الأعمال المرتبطة WORK RELATED
The Long-Horizon Task Mirage? Diagnosing Where and Why Agentic Systems
Break
arXiv:2604.11978
cross-domain horizon degradation and trajectory-grounded failure attribution.
Long-Horizon Agent Trajectory Attribution
arXiv:2608.06909
fine-grained attribution chains and component-level leave-one-out perturbation.
الدمج لا يتم بالتشابه الاسمي؛ نستخرج ما يغير object formal أو experiment. إذا لم يغير المرجع boundary أو method يبقى background.
2/4 الأعمال المرتبطة WORK RELATED
Trustworthy agents in practice
Anthropic, 2026-04-09
human control, values, secure interaction, transparency and privacy in agent deployments.
How we contain Claude across products
Anthropic, 2026-05-25
containment and blast-radius engineering; permission-fatigue limits human gating.
الدمج لا يتم بالتشابه الاسمي؛ نستخرج ما يغير object formal أو experiment. إذا لم يغير المرجع boundary أو method يبقى background.
3/4 الأعمال المرتبطة WORK RELATED
Securing the future of AI agents / AI Control Roadmap
Google DeepMind, 2026-06-18
detection/prevention/response, supervisors and control metrics for increasingly capable
agents.
NIST Concept Paper: Software and AI Agent Identity and Authorization
NIST NCCoE, 2026-02-05
identity and authorization controls for software/AI agents accessing diverse tools and data.
الدمج لا يتم بالتشابه الاسمي؛ نستخرج ما يغير object formal أو experiment. إذا لم يغير المرجع boundary أو method يبقى background.
4
<PARSED TEXT FOR PAGE: 5 / 25>
4/4 الأعمال المرتبطة WORK RELATED
Model Context Protocol Specification 2026-07-28
MCP 2026-07-28
stateless core, tasks extension, authorization hardening and deprecation lifecycle.
A2A Protocol v1.0 / Agent2Agent Specification
A2A 1.0.0
interoperable discovery, task management and communication between opaque agent
systems.
الدمج لا يتم بالتشابه الاسمي؛ نستخرج ما يغير object formal أو experiment. إذا لم يغير المرجع boundary أو method يبقى background.
1/3 السلالة الداخلية LINEAGE MUT
13-06M — نظرية مساعف لطوبولوجيا الوصول والالتزام
كل فعل يغير graph reachability للنظام. reversibility ليست label ثنائي؛ هي بنية لعدد ونوعية حالات
الرجوع والبدائل المتبقية بعد الالتزام، تحت الموارد والصلاحيات والزمن.
CommitCost(a)=LossReachability + λ·RollbackDistance + μ·AuthorityExposure + ν·TemporalLockIn
13-07M — نظرية مساعف للوكالة الحاملة للبرهان
كل فعل مادي عالي الأثر يحمل حزمة برهان دنيا: authority، preconditions، expected state delta،
uncertainty، observation plan، rollback/compensation، verifier independence، وsuccess
criterion. الفعل غير القادر على حملها لا يترقى إلى execution autonomous.
(ActionAdmissible = Auth ∧ Pre ∧ ObsPlan ∧ RiskGate ∧ (Rollback ∨ ExplicitIrreversibleApproval
Part XIV لا يستبدل هذه parents؛ يحدد أين بقي closure ناقصاً ثم يضيف specific-theory variable state أو law consistency أو

.mapping evaluation
5
<PARSED TEXT FOR PAGE: 6 / 25>
2/3 السلالة الداخلية LINEAGE MUT
13-09M — نظرية مساعف للوكالة المقاومة لانهيار الأفق
وكالة طويلة الأفق تقاس بقدرتها على المحافظة على obligations/evidence/state/authority/goal مع زيادة horizon وdepth dependency، وبسرعة اكتشاف وتصحيح الانحراف. الأداء النهائي وحده يخفي تراكم الأخطاء.
(HRA = ∫_0^H [TaskIntegrity(h) × StateFidelity(h) × Recovery(h)] dh / (1 + CascadeRisk
13-15M — نظرية مساعف لإسناد السلطة وكشف الاستحواذ العكسي
كل قرار impact-high يحمل graph attribution للسلطات والإشارات التي أثرت فيه. الاستحواذ يحدث عندما يحصل مصدر منخفض السلطة أو غير موثوق على influence causal يفوق tier المسموح، حتى لو بدا القرار النهائي معقولاً .
(CaptureRatio = Σ(c_i for unauthorized/low-tier) / Σ(c_j for authorized sources
XIV Part لا يستبدل هذه parents، يحدد أين بقي closure ناقصاً ثم يضيف variable state specific-theory أو law consistency أو .mapping evaluation
3/3 السلالة الداخلية LINEAGE MUT
13-19M — نظرية مساعف للإغلاق السببي بين النظرية والنواة
الإغلاق السببي يتحقق عندما يؤدي تدخل معرف في construct النظري إلى تغيير measurable في kernel mechanism ثم trajectory ثم outcome evidence/world، مع attribution قابلة لإعادة التشغيل. غياب أي حلقة يكسر claim التفسيري.
ClosureStrength = Causal(T→K) × Causal(K→τ) × Causal(τ→W) × EvidenceReproducibility
XIV Part لا يستبدل هذه parents، يحدد أين بقي closure ناقصاً ثم يضيف variable state specific-theory أو law consistency أو .mapping evaluation
6
<PARSED TEXT FOR PAGE: 7 / 25>
M14-04 PAPER-XIV-02 — النظري الاشتقاق
نظرية مساعف للمسار الحامل للبرهان
invariant ledger، pending obligations، authority budget، evidence continuity، and. ، كاف غير شرط محلية الخطوة صحة
كل trajectory تحمل طويلة proof state متراكماً : rollback anchors
تعتمد أيضاً على أثرها على برهان المسار. admissibility
TrajectoryAdmissible_t = ActionProof_t × InvariantRetention_t × ObligationConsistency_t × AuthorityBudget_t × RecoveryReachability_t
الفجوة: Agency Carrying-Proof يضبط action عالي الأثر، للكن failure قد ينشأ من تسلسل أفعال كل منها صالح منفرداً بينما trajectory الكاملة تنتهك الهدف أو السلطة أو budget.
يمثل هذا الجسم النظري انتقالاً من 13-parent theories 13-07M · 13-09M · 19M إلى quantity جديدة ذات دور قبل النتيجة. الاستقلال لا يُقاس بوجود صيغة جديدة، بل بقدرة intervention على المتغير المركزي على تغيير prediction أو signature failure مع بقاء أقرب confounders ثابتة. إذا كان parent أبسط ينتج نفس المسار، يجب دمج النظرية.
Objects
TrajectoryProofState · InvariantLedger · ObligationStack · AuthorityBudget · RecoveryAnchor
Predictions
يكشف failures تراها لا per-action policies.
تحسن safety horizon-long أكثر من زيادة confirmations.
يخفض authorization drift عبر turns +100.
Boundary: ليس verification formal كاملاً ؛ هو obligations proof قابلة للتتبع.
•
•
•
7
<PARSED TEXT FOR PAGE: 8 / 25>
M14-04 PAPER-XIV-02 — والخوارزمية الآلية
ابدأ proof state من goal/authority.
كل action تحقق action proof + trajectory invariants.

.budgets/obligations حدّث
.readback/نفذ
اكتشف cumulative violations.
checkpoint proof state.
rollback أو replan عند threshold.
تنفذ الآلية تخدمة قابلة للتعطيل. تدخل STATE-PRE موثقة، تولد candidates، تمر عبر gate authority/evidence حسب طبيعتها، ثم تنتج transition ولـ residual. أي residual كبير لا يعالج بإعادة المحاولة فقط؛ يصبح evidence
على فشل model أو tool أو assumption. هذه القاعدة تمنع تحويل policy retry إلى قناع للفشل النظري.
Failure tests
per-action proof يساوي الأداء.
proof state دون يتضخم predictive value.
invariant ledger لا يتنبأ trajectory failure.
Interactions
مع التفاعلات تختبر parents M13-07 · M13-09 · M13-19 وsibling theories عبر sequential/
. اعتباطياً credit من توزيع بدلاً تسجل uncertainty attribution .ablations factorial
كل object state يحمل uncertainty/provenance/version. لا يصعد إلى layer procedural/memory دون rollback .receipt جزء من algorithm وليس مرحلة طارئة خارج التصميم.
M14-04 PAPER-XIV-02 — والترقية الاختبار
HORIZON/UltraHorizon-like tasks :Decisive experiment مع cumulative traps؛ compare per-action gates vs PCTT.
Requirement Control
Matched budget نفس model/tools/context/retries/time الإمكان قدر.
Sham scaffold نفس overhead دون theory-specific transition أثر لعزل compute والتنظيم.
حالات يجب ألا تُفعل فيها النظرية. activation Negative
domain/model/tool family Transfer مع ثانية preregistered metric.
3N يحتاج trace؛ 4N يحتاج gain causal مع CI وregression critical no؛ 5N يحتاج replication مستقلة. الفشل يُحفظ ويضيق claim أو يدمج النظرية. لا توجد حماية لأن النظرية من corpus الداخلي.
.Promote only when causal prediction survives ablation, transfer and guardrails
1.
2.
3.
4.
5.
6.
7.
•
•
•
8
<PARSED TEXT FOR PAGE: 9 / 25>
M14-05 PAPER-XIV-02 — النظري الاشتقاق
نظرية مساعف لهندسة أفق الالتزام
الالتزام له هندسة زمنية حيث set rollback future ينكمش بمرور الوقت وتكاثر dependents. يجب قياس horizon commitment والوقت المتبقي قبل أن يصبح التراجع مكلفاً أو مستحيلاً .
CHG(a,t)=OptionLoss_t + DependencyGrowth_t + RollbackCost_t + DeadlineLockIn_t; CriticalTime=min t: CHG>τ
الفجوة: topology Reachability تقيس فقد الخيارات، للكن commitments لها زمن: بعض القرارات قليلة الكلفة الآن وتخلق in-lock بعد دقائق/مراحل عندما تتراكم dependencies.
يمثل هذا الجسم النظري انتقالاً من parent-13 theories 06M-13 · 09M-13 · 11M إلى quantity جديدة ذات دور قبل النتيجة.
الاستقلال لا يُقاس بوجود صيغة جديدة، بل بقدرة intervention على المتغير المركزي على تغيير prediction أو signature failure مع بقاء أقرب confounders ثابتة. إذا كان parent أبسط ينتج نفس المسار، يجب دمج النظرية.
Objects
CommitmentHorizon · DependencyCone · CriticalRollbackTime · OptionSet · LockInEvent
Predictions
earlier checkpoints تمنع costly reversals.

عامل الزمن يفسر failures لا تفسرها score risk لحظية.
agent يتصرف differently مع نفس action عند horizons مختلفة.
Boundary: تعتمد تقريبية هندسة knowledge of dependencies.
•
•
•
9
<PARSED TEXT FOR PAGE: 10 / 25>
M14-05 PAPER-XIV-02 — والخوارزمية الآلية
estimate dependency cone بعد action.
forecast rollback cost عبر horizon.
حدد critical rollback time.
ضع checkpoints قبلها.
أعد تقييم evidence قبل عبور in-lock.
سجل irreversible transition receipt.
تنفذ الآلية نكدمة قابلة للتعطيل. تدخل STATE-PRE موثقة، تولد candidates، تمر عبر gate authority/evidence، حسب طبيعتها، ثم تنتج transition ولا residual. أي residual كبير لا يعالج بإعادة المحاولة فقط؛ يصبح evidence
على فشل model أو tool أو assumption. هذه القاعدة تمنع تحويل policy retry إلى قناع للفشل النظري.
Failure tests
static reachability score يكفي.
estimate time-critical غير قابل للمعايرة.
checkpoints لا تخفض reversal cost.
Interactions
مع التفاعلات تختبر parents M13-06 · M13-09 · M13-11 وsibling theories عبر sequential/
. اعتباطياً credit من توزيع بدلاً uncertainty attribution تسجل ablations factorial.
كل object state يحمل uncertainty/provenance/version. لا يصعد إلى layer procedural/memory دون rollback. receipt جزء من algorithm وليس مرحلة طارئة خارج التصميم.
M14-05 PAPER-XIV-02 — والترقية الاختبار
code migrations/cloud deployments/research pipelines :Decisive experiment مع dependencies متراكمة.
Requirement Control
Matched budget نفس model/tools/context/retries/time الإمكان قدر.
Sham scaffold نفس overhead دون theory-specific transition أثر لعزل compute والتنظيم.
activation Negative حالات يجب ألا تُفعل فيها النظرية.
domain/model/tool family Transfer مع ثانية preregistered metric.
3N يحتاج trace؛ 4N يحتاج gain causal مع CI وno regression critical؛ 5N يحتاج replication مستقلة. الفشل يُحفظ ويضيق claim أو يدمج النظرية. لا توجد حماية لأن النظرية من corpus الداخلي.
.Promote only when causal prediction survives ablation, transfer and guardrails
1.
2.
3.
4.
5.
6.
•
•
•
10
<PARSED TEXT FOR PAGE: 11 / 25>
M14-06 PAPER-XIV-02 — النظري الاشتقاق
نظرية مساعف لحفظ السلطة عبر التفويض
conservation لا يكتسب rights من أوسع intersection بين user authority، delegator scope، protocol contract، destination المفوضة السلطة constraint: downstream agent/tool
policy. أي غموض يخفض scope ولا يرفعه.

Authority_child = UserAuthority ∩ DelegatorGrant ∩ ProtocolScope ∩ DestinationPolicy ∩ TemporalValidity
الفجوة: MCP/A2A يسهّلان delegation agent/tool، للكن interoperability لا يضمن أن الصلاحية لا تتوسع أو تتشوه عبر handoffs.
يمثل هذا الجسم النظري انتقالاً من parent-13 theories 13-07M · 13-15M · 20M إلى quantity جديدة ذات دور قبل النتيجة.
الاستقلال لا يُقاس بوجود صيغة جديدة، بل بقدرة intervention على المتغير المركزي على تغيير prediction أو signature failure مع بقاء أقرب confounders ثابتة. إذا كان parent أبسط ينتج نفس المسار، يجب دمج النظرية.
Objects
DelegationGrant · AuthorityEnvelope · AgentIdentity · ScopeIntersection · RevocationReceipt
Predictions
يقل privilege amplification في multi-agent chains.
protocol capability lists تصبح لا permission تلقائياً.
revocation propagation تصبح measurable reliability dimension.
Boundary: يبني فوق IAM ولا يستبدله.
•
•
•
11
<PARSED TEXT FOR PAGE: 12 / 25>
M14-06 PAPER-XIV-02 — والخوارزمية الآلية
resolve agent/tool identity.
compile authority envelope.
intersect scopes at handoff.
bind expiry/task/object.
execute within envelope.
propagate revocation.
audit causal influence if low-tier content changes action.
تنفذ الآلية نكخدمة قابلة للتعطيل. تدخل STATE-PRE موثقة، تولد candidates، تمر عبر gate authority/evidence حسب طبيعتها، ثم تنتج transition وresidual. أي residual كبير لا يعالج بإعادة المحاولة فقط؛ يصبح evidence على فشل model أو tool أو assumption. هذه القاعدة تمنع تحويل policy retry إلى قناع للفشل النظري.
Failure tests
standard IAM delegation يحقق نفس guarantees بلا semantic layer.
scope intersections تمنع لا injected escalation.
identity uncertainty يرتبط لا incidents.
Interactions
مع التفاعلات تختبر parents M13-07 · M13-15 · M13-20وsibling theories عبر sequential/
اعتباطياً. تسجل uncertainty attribution .ablations factorial بدلاً من توزيع credit
جزء من rollback .receipt دون layer procedural/memory. لا يصعد إلى uncertainty/provenance/version يحمل object state كل
وليس مرحلة طارئة خارج التصميم. algorithm
M14-06 PAPER-XIV-02 — والترقية الاختبار
A2A/MCP mock organizations :Decisive experiment مع nested delegation وprompt injection.
Requirement Control
Matched budget نفس model/tools/context/retries/time الإمكان قدر.
Sham scaffold نفس overhead دون theory-specific transition أثر لعزل compute والتنظيم.
حالات Negative activation يجب ألا تُفعل فيها النظرية.
domain/model/tool family Transfer مع ثانية preregistered metric.
3N يحتاج trace؛ 4N يحتاج gain causal مع CI وregression critical no؛ 5N يحتاج replication مستقلة. الفشل يُحفظ ويضيق claim أو يدمج النظرية. لا توجد حماية لأن النظرية من corpus الداخلي.
.Promote only when causal prediction survives ablation, transfer and guardrails
1.
2.
3.
4.
5.
6.

7.
•
•
•
12
<PARSED TEXT FOR PAGE: 13 / 25>
النموذج الموحد للحالة MODEL FORMAL
TrajectoryAdmissible_t = ActionProof_t × InvariantRetention_t × ObligationConsistency_t × AuthorityBudget_t ×
+ RecoveryReachability_t ; CHG(a,t)=OptionLoss_t
DependencyGrowth_t + RollbackCost_t + DeadlineLockIn_t; CriticalTime=min t: CHG>τ ; Authority_child =
∩ UserAuthority ∩ DelegatorGrant ∩ ProtocolScope
DestinationPolicy ∩ TemporalValidity
تعامل الصيغ كمجموعة constraints لا scalar واحد. تُبنى vector state من objects النظرية المختارة، وتُحدد variables hard التي لا يجوز تعويضها بـutility: authority، unknowns critical، approvals irreversible،
provenance evidence. optimization يعمل بعد ذلك داخل region المسموح.
three-level state: epistemic (belief/unknown/evidence)، operational. آخر مستوى فيً ضمنيا يعالج لا المستويات أحد في الخطأ.
نستخدم ((plans/tools/memory/protocol state)، world (observed effects
قابلية التعيين والسببية IDENTIFIABILITY
لكل construct تدخل minimal. نمنع interventions التي تغير length prompt أو retries أو simultaneously model. عند صعوبة العزل نستخدم design factorial ونقدر interaction terms. causal claim لا ينتج من
correlation بين theory score وال success؛ تتغير أن يجب prediction بعد (do(variable في وتظهر trajectory/world effect.
Causal closure = Theory intervention → Kernel delta → Trajectory delta → World/Evidence delta
13
<PARSED TEXT FOR PAGE: 14 / 25>
النماذج المضادة COUNTERMODELS
Baseline أبسط يحقق الأداء نفسه: contribution تندمج.
Sham scaffold نفس يحقق gain: السبب compute/organization لا theory.
Metric دون تتحسن world/evidence gain: metric gaming.
نتيجة محلية لا تنتقل: claim يُضيق.
Safety/authority regression: rollback كان مهما gain.
Judge مع مشترك candidate يفسر التحسن: evaluator-dependence debt.
الديناميات بين المقاييس OBJECTIVE-MULTI
verified success، calibration، unknown retention، latency، tokens/tools، human oversight، recovery، irreversible، errors، critical violations يشمل vector. accuracy في الأداء نخختزل لا
reproduction. report يعرض Pareto frontier الأهداف تتعارض عندما. hard governance constraints تبقى vetoes.
.Utility optimization is permitted only inside the hard-constraint feasible region
القانون الإغلاقي للورقة LAW CLOSURE
الإغلاق المقترح في هذه الورقة: سلامة الفعل لا تكفي إذا كانت trajectory كاملة تنتهك الهدف أو السلطة. نقترح state proof متراكماً للمسار، protocols/tools/agents. هندسة زمنية للالتزام، وقانون حفظ للسلطة أثناء التفويض عبر
لا تُعتبر الحلقة مغلقة إذا توقفت عند reasoning أو simulator أو verifier internal. يجب وجود readback world/source مناسب لنوع الادعاء، attribution كافٍ ، ثم writeback يحفظ result negative والunknowns.
يحتاج تكرار الحلقة بعد claim generality transformation.
1.
2.
3.
4.
5.
6.
14
<PARSED TEXT FOR PAGE: 15 / 25>
الخوارزمية والتنفيذ ALGORITHM 1/5
goal/authority/evidence/؛ افصل Unknown؛ اربط versions. ليس الناتج بل prompt object graph قابلا للتسلسل وإعادة البناء.
Algorithm A: BuildTypedState. اجمع world/memory variables

تقترح hypotheses/actions دون commit. كل candidate يحمل expected delta، cost، proof obligations وfailure prediction.
Algorithm B: GenerateCandidates. reasoning/search/agents
2/5 الخوارزمية والتنفيذ ALGORITHM
اختر أصغر فعل كامل قابل للرجوع يميز .gates trajectory/reversibility/evidence/authority طبق .GateAndAct :C Algorithm
أقوى. proof أو approval يتطلب impact high عند .utility أو يحقق hypotheses
افصل tool return عن world verification. قارن expected/observed state، واكتب residual مع attribution uncertainty.
Algorithm D: ReadbackAndResidual.
3/5 الخوارزمية والتنفيذ ALGORITHM
؛ semantic/procedural promotion ينتظر recurrence/causal eligibility/transfer. contradiction لا overwrite، يفتح branch.
Algorithm E: LearnAndWriteback. immediate residue يذهب episodic
Algorithm F: Replicate. أعد run في model/domain/tool/horizon مختلف، مع harness fingerprint وsame claim boundary.
4/5 الخوارزمية والتنفيذ ALGORITHM
يملك switchable services؛ Evidence Lab يملك run manifests؛ Evolution Lab يملك candidates/rollback. لا service حق تملك
Architecture: MUT يملك theory/lineage؛ MUG يختار vessel؛ MOUSSAIXT
وحدها. promotion
15
<PARSED TEXT FOR PAGE: 16 / 25>
5/5 الخوارزمية والتنفيذ ALGORITHM
causal attribution أسهل، ثم browser/code environments، ثم long-running/multi-agent. لا تنقل إلى weights قبل أن يصبح
Implementation sequencing: ابدأ mock/simulator deterministic حيث
مستقراً وقابلاً للقياس. mechanism external
1/4 التجربة EXPERIMENT
HORIZON-like long tasks مع cumulative authorization traps.
preregistered hypothesis; matched budgets; repeated seeds; independent/alternative evaluator where possible Design.
theory-specific metric + verified task/world effect + calibration + cost Primary outcomes.
conditions where mechanism must remain inactive or abstain Negative cases.
N4 only if causal gain and no critical regression; N5 after transfer/replication Promotion.
2/4 التجربة EXPERIMENT
Code/cloud migration tasks لقياس Critical Rollback Time.
preregistered hypothesis; matched budgets; repeated seeds; independent/alternative evaluator where possible Design.
theory-specific metric + verified task/world effect + calibration + cost Primary outcomes.
conditions where mechanism must remain inactive or abstain Negative cases.
N4 only if causal gain and no critical regression; N5 after transfer/replication Promotion.
16
<PARSED TEXT FOR PAGE: 17 / 25>
3/4 التجربة EXPERIMENT
Nested A2A/MCP delegation مع injected low-tier instructions.
preregistered hypothesis; matched budgets; repeated seeds; independent/alternative evaluator where possible Design.
theory-specific metric + verified task/world effect + calibration + cost Primary outcomes.
conditions where mechanism must remain inactive or abstain Negative cases.
N4 only if causal gain and no critical regression; N5 after transfer/replication Promotion.
4/4 التجربة EXPERIMENT
Containment vs proof-state ablation under matched permissions.
preregistered hypothesis; matched budgets; repeated seeds; independent/alternative evaluator where possible Design.
theory-specific metric + verified task/world effect + calibration + cost Primary outcomes.
conditions where mechanism must remain inactive or abstain Negative cases.
N4 only if causal gain and no critical regression; N5 after transfer/replication Promotion.
مصفوفة الاجتثاث ABLATION
Purpose Mechanism Run
candidate effect ON A
nearest baseline OFF B
same overhead without theory transition SHAM C
tests directionality and metric gaming PERTURBED D

A≈C،. إذا A>others وللكن prediction pattern لا يطابق النظرية، implementation may help but theory is not supported.
إذا theory-specific attribution fails
17
<PARSED TEXT FOR PAGE: 18 / 25>
مخاطر القياس والفشل
FAILURE MODES
LLM judge dependence بين والتسرب candidate/evaluator.
harness changes masquerading as capability.
observation proxies that do not measure world effect.
retry loops hiding systematic failure.
dataset/task leakage and post-hoc metric selection.
under-accounted human work, tool cost or permissions.
theory names surviving after mechanism equivalence is shown.
الحدود العلمية LIMITATIONS
هذه الورقة تقدم نظرية وبرنامجاً تجريبياً ، لا نتائج 5E4/E. بعض الصيغ heuristic أو functionals تحتاج 2026 calibration. literature سريع
التغير وبعض المراجع preprints أو research vendor. corpus MOUSSAIF نفسه
واسع وقد يحتوي صيغاً تاريخية متداخلة؛ لذلك تُستخدم boundaries lineage/canonical ولا تدعى novelty العالمية قبل review art-
.prior dedicated
Stack IMPLEMENTATION STACK التنفيذ
Responsibility Layer
theory objects, parents, contradictions, evidence status MUT
Context/HyperKB/Skill/Loop/Adapter/Weights placement MUG
services, agents, tools, feature flags, telemetry MOUSSAIXT
manifests, replay, metrics, CI, negative results Evidence Lab
branch, canary, rollback, external evaluator Evolution Lab
•
•
•
•
•
•
•
18
<PARSED TEXT FOR PAGE: 19 / 25>
MUG TRAINING PLACEMENT والتدريب
ابدأ mechanism/knowledge خارجياً للحفاظ على editability. انقل إلى HyperLoop/HyperSkill بعد use repeated. عندما Adapter
يكون السلوك specific-protocol/provider. weights فقط عندما يثبت transfer
والretention ويصبح الاستدعاء الخارجي bottleneck. التدريب يعلن نوعه بدقة ولا يسمى كل تحسين fine-tuning.
Reproducibility package REPRODUCIBILITY
commit/version، model/harness/tool identifiers، environment image، prompts/context pointers، budget، seed،
ترفق تجربة كل: evaluator، raw trajectory، world readback، metrics، failure
traces، script analysis، verdict. إذا تعذر نشر جزء لأسباب أمنية يسجل exclusion وسببها.
.Trace or it did not happen. Replay where possible; declare irreproducible dependencies where not
الخلاصة CONCLUSION
سلامة الفعل لا تكفي إذا كانت trajectory كاملة تنتهك الهدف أو السلطة. نقترح state proof متراكماً للمسار، هندسة زمنية للالتزام، وقانون
حفظ للسلطة أثناء التفويض عبر protocols/tools/agents.
القيمة النهائية لهذه الورقة لن تكون في بقائها كنظرية، بل في قدرتها على إنتاج experiment يفرق بينها وبين baseline/parent. إذا لم تنجح، يجب
أن تُدمج أو تُضيق. إذا نجحت محلياً فقط، تبقى mechanism local. واذا نجحت عبر
transformations مع evidence مستقلة، يمكن أن تصبح جزءاً أقوى من MUG/MUT closure.
19
<PARSED TEXT FOR PAGE: 20 / 25>
المراجع ومصادر السلالة REFERENCES
Role Reference External source

The Long-Horizon Task Mirage? Diagnosing Where and Why Agentic Systems
Break
cross-domain horizon degradation and trajectory-grounded failure attribution arXiv:2604.11978.
fine-grained attribution chains and component-level leave-one-out perturbation arXiv:2608.06909 Long-Horizon.
Agent Trajectory Attribution
human control, values, secure interaction, transparency and privacy in agent Anthropic, 2026-04-09 Trustworthy agents in practice
deployments.
containment and blast-radius engineering; permission-fatigue limits human Anthropic, 2026-05-25 How we contain Claude across products
gating.
detection/prevention/response, supervisors and control metrics for increasingly Google DeepMind, 2026-06-18 Securing the future of AI agents / AI Control Roadmap
capable agents.
identity and authorization controls for software/AI agents accessing diverse NIST NCCoE, 2026-02-05 NIST Concept Paper: Software and AI Agent Identity and Authorization
tools and data.
stateless core, tasks extension, authorization hardening and deprecation lifecycle MCP 2026-07-28 Model Context. Protocol Specification 2026-07-28
interoperable discovery, task management and communication between opaque A2A 1.0.0 A2A Protocol v1.0 / Agent2Agent Specification
agent systems.
Role Theory Internal
parent XIII Part نظرية مساعف لطوبولوجيا الوصول والالتزام 13-06M
parent XIII Part نظرية مساعف للوكالة الحاملة للبرهان 13-07M
parent XIII Part نظرية مساعف للوكالة المقاومة لانهيار الأفق 13-09M
parent XIII Part نظرية مساعف لإسناد السلطة وكشف الاستحواذ العكسي 13-15M
parent XIII Part نظرية مساعف للإغلاق السببي بين النظرية والنواة 13-19M
Internal source corpus also includes Parts X-XIV, MUT-AGI-02, MUG/MOUSSAIXT framework, Part XI memory/ continuity, Part XII paper integration, and the isolated research-paper registry. External sources are design pressure/ prior art, not validation of
MOUSSAIF.
20
<PARSED TEXT FOR PAGE: 21 / 25>
قائمة تحقق المراجعة قبل النشر CHECKLIST REVIEW
هل كل claim أوسع من evidence تم تضييقه؟
هل results external منفصلة عن hypotheses internal؟
هل equations تحمل definitions operational؟
هل baseline متطابق الموارد؟
هل activation negative موجود؟
هل world/evidence readback مستقل؟
هل dependence evaluator موثق؟
هل failure/rollback path موجود؟
هل claim transfer له replicate؟
هل theory يمكن reject/merge دون كسر الهوية المشروع؟
A الملحق A APPENDIX — قاموس الرموز والمتغيرات التشغيلية
تستعمل الورقة مجموعة متغيرات مشتركة للكنها لا تفترض أن كلها observable مباشرة: CommitmentHorizon · RecoveryAnchor · AuthorityBudget · ObligationStack · InvariantLedger · TrajectoryProofState ·
DependencyCone · CriticalRollbackTime · OptionSet · LockInEvent · DelegationGrant · AuthorityEnvelope ·. AgentIdentity · ScopeIntersection · RevocationReceipt
TrajectoryAdmissible_t = ActionProof_t × InvariantRetention_t × ObligationConsistency_t × AuthorityBudget_t × + RecoveryReachability_t ; CHG(a,t)=OptionLoss_t
DependencyGrowth_t + RollbackCost_t + DeadlineLockIn_t; CriticalTime=min t: CHG>τ ; Authority_child =

∩ UserAuthority ∩ DelegatorGrant ∩ ProtocolScope
DestinationPolicy ∩ TemporalValidity
لكل متغير خمس سمات إلزامية: domain، estimator، uncertainty، provenance، interventionability. المتغير الذي يبقى حكماً لغوياً بلا estimator لا يدخل equations causal. وإذا كان based-model estimator يسجل
model/version/calibration. أما hard governance constraints تدخل فلا tradeoff scalar.
Rule Examples Class
readback required observed state, event, effect World
cannot silently vanish belief, unknown, residual Epistemic
scope/expiry/revocation authority, commitment, delegation Agency
must not rewrite runtime facts score, CI, cost Evaluation
1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
21
<PARSED TEXT FOR PAGE: 22 / 25>
APPENDIX B الملحق B — مقترحات واختبارات اتساق
مناسبة. perturbation عند outcome قبل prediction السببي يجب أن يغير sensitivity Intervention: construct
.conflictك يظهر conflict ،receipt بلا uncertainty مستقلة جديدة لا تمحو monotonicity Evidence: evidence
. تلقائياً authority الأعلى لا توسع autonomy Bounded: capability
لا النص الحرفي. tolerance وسببها ضمن class verdict يعيد verdict Replayable: replay
generality claim :Transfer scope يتجاوز لا transformation family المختبرة.
هذه hypotheses propositions تشغيلية لا theorems نهائية. وظيفتها أن تجعل التجربة قادرة على الرفض وأن تمنع تحويل score واحد إلى دعم عام.
APPENDIX C الملحق C — الخوارزمية المرجعية والتكلفة
materialize typed state → expose unknown/authority/budget → generate candidates → estimate expected epistemic/ → world delta → compile proof obligations :Reference procedure
choose smallest complete reversible intervention → execute/observe → readback → residual/attribution → memory/. theory update
Model + Tool + MemoryRead + MemoryWrite + Verification + Human + Waiting/ تخفض قد active compute زيادة مع
يساوي لا التعقيد LLM calls فقط. نميز Opportunity. long-running agents
time elapsed؛ لذلك يبلغان منفصلين.
TotalCost = C_model + C_tool + C_memory + C_verify + C_human + C_wait
أي optimization تعلن أي cost خفضته وأي risk قد رفعته. tokens وحدها ليست efficiency.
1.
2.
3.
4.
5.
22
<PARSED TEXT FOR PAGE: 23 / 25>
APPENDIX D الملحق D — بناء benchmark تشخيصي
الورقة لهذه الأساسية التجارب: model/tool/domain. ثم transfer tasks تغير drift/recovery، تختبر interactions، long-horizon tasks تختبر construct، meso tasks تختبر. كل micro tasks تعزل من suite بنُى ت
cumulative authorization لقياس Critical Rollback Time.; Nested A2A/MCP delegation مع injected low-tier instructions.; HORIZON-like long tasks مع traps.; Code/cloud migration tasks
Containment vs proof-state ablation under matched permissions..

missing observation، stale memory، malicious tool text، permission change، delayed event،. لكل perturbation expected
لكل task perturbations قبل مسماة الاختبار: tool failure، distribution shift
failure signature. يمنع هذا post-hoc storytelling.
Use Split
calibration only Dev
primary causal verdict Held-out
boundary discovery Adversarial
N5/generalization Transfer
E الملحق E APPENDIX — خطة التحليل الإحصائي
mean/ بحسب metric، CI، عدد seeds/tasks، وتوزيع failures. في paired tasks نستخدم paired analysis. costs تعرض كsuccess-
secondary diagnostics. نبلغ median وصغيرا primary outcome نحدد التشغيل قبل
at-budget أو Pareto front.
مع تعدد النظريات نستخدم gate level-family ثم confirmation على out-held بدلاً من انتقاء أفضل .value-p تبقى findings exploratory
وحدها، significance أهم من size effect .replication حتى exploratory
والنتيجة السلبية ذات CI ضيق قد تكون علمياً أثمن من gain صغير هش.

F الملحق F APPENDIX — تهديدات الصلاحية والأمن
proxy :validity Construct لا يمثل المفهوم.
Internal validity: harness/compute/retries confound.
External validity: transfer غير مختبر.
Evaluator dependence: judge shares roots.
Security externality: tool/agent content manipulates policy.
Authority leakage: delegation expands privilege.
Memory contamination: false lesson becomes canonical.
Publication bias: failed runs disappear.
instrument/experiment: sham control، alternate judge، prompt-injection suite، authority envelope، writeback.
كل threat يقابله eligibility، negative-result registry
G الملحق G APPENDIX — Crosswalk إلى MUT/MUG/MOUSSAIXT
M14-04 · M14-05 · M14-06. parents:. في MUT تحفظ definitions/parents/contradictions؛ في MUG يحدد vessel؛ في
المركزية النظريات: M13-06 · M13-07 · M13-09 · M13-15 · M13-19
MOUSSAIXT تنفذ services؛ في Evidence Lab تربط runs؛ في Evolution Lab تحكم promotion/rollback.
Mandatory object Layer
definition,parent,falsifier,status MUT
vessel,reversibility,transfer MUG
service,interface,telemetry,failure MOUSSAIXT
baseline,manifest,result,CI Evidence
branch,evaluator debt,rollback Evolution
•
•
•
•
•
•
•
•

H الملحق H APPENDIX — برنامج الأسئلة المفتوحة
ما أقل representation state تكفي دون modeling-over؟
ما metrics التي تقاس من state world مباشرة بدل judges LLM؟
متى يصبح overhead verification/memory أكبر من فائدته؟
كيف نفصل base-model capacity عن system/harness contribution؟

أي signatures failure تنتقل عبر domains؟
كيف نقيم improvement-self عندما evaluator نفسه يتطور؟
كيف نحسب oversight human لـresource لا جزء مجاني؟
أي theories يجب دمجها بعد أول 4N3/N cycle؟
القائمة جزء من contribution لأنها تمنع closure زائفاً : البرنامج القوي يحدد ما لا يعرفه وما experiment الذي يغير رأيه.

1.
2.
3.
4.
5.
6.
7.
8.