RECONSTRUCTED_FROM_INDEXED_TEXT - نسخة أرشيفية معاد بناؤها من النص المفهرس

<PARSED TEXT FOR PAGE: 1 / 24>
PAPER-XIV-04 · MOUSSAIF AGI RESEARCH · 2026
الثقة في التفويض سيادة البروتوكول والتنظيم متعدد الوكلاء: حمل المحاذاة ومعايرة
Protocol-Sovereign Multi-Agent Intelligence: Organizational Alignment Load and Calibrated Delegation Trust
التعاون بين agents يرفع capability للكنه قد يرفع misalignment والسطح الهجومي. نقترح Load Alignment Organizational، طبقة فوق Sovereignty Semantic-Protocol
A2A/MCP، وثقة تفويض مشروطة بالسياق لا scalar reputation.
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 / 24>
ABSTRACT الملخص والادعاء المركزي
التعاون بين agents يرفع capability للكنه قد يرفع misalignment والسطح الهجومي. نقترح Load Alignment Organizational، طبقة فوق Sovereignty Semantic-Protocol، وثقة تفويض مشروطة بالسياق لا A2A/MCP reputation scalar.
تبني الورقة على XII Part الذي فصل presence عن integration، وXIII Part الذي صاغ closure causal world-open verified، ثم تستخدم XIV Part لإغلاق طبقة محددة بآليات قابلة للتعطيل والاختبار. الأدبيات الخارجية تُستخدم كدليل أن MOUSSAIF تُستخدم لـart prior وتصميم pressure failure/benchmark، ولا تحقق نتائجها.
Contributions
قياس emergent organizational risk من ناشئ لحمل delegation/incentives/asymmetry/authority ambiguity.
فصل transport interoperability عن semantic sovereignty للgoal/authority/state/evidence/rollback.
Trust posterior conditional على task/horizon/tools/authority مع failure attribution.
تصميم organizations كـgraphs قابلة لإعادة الهيكلة عندما يتجاوز الحمل capacity governance.
Scientific value = distinct causal hypothesis × falsifiability × reproducibility × transfer
1.
2.
3.
4.
2
<PARSED TEXT FOR PAGE: 3 / 24>
CONTRIBUTIONS المساهمات وحدود الأسبقية
Status Contribution
new synthesis within MOUSSAIF; requires literature novelty audit and experiments .delegation/incentives/asymmetry/ قياس emergent organizational risk من ناشئ لحمل authority ambiguity
new synthesis within MOUSSAIF; requires literature novelty audit and experiments .goal/authority/state/evidence/ فصل transport interoperability عن semantic sovereignty للrollback
task/ مع new synthesis within MOUSSAIF; requires literature novelty audit and experiments .failure attribution Trust posterior conditional على horizon/tools/authority
الحمل يتجاوز new synthesis within MOUSSAIF; requires literature novelty audit and experiments .governance capacity تصميم organizations كgraphs قابلة لإعادة عندما الهيكلة
novelty .protocols/control/ لا تدعي هذه الورقة أن المصطلحات أو كل المكونات بلا سوابق عالمية؛ المجتمع العلمي يملك خطوطاً قوية في benchmarks/memory/models world/reasoning agentic المقترحة تقع في تركيبها إلى
closure law وفي formal distinctions داخل المشتركة والاختبارات MUT/MUG/MOUSSAIXT.
DISCIPLINE EVIDENCE منهج المصدر والدليل
internal source-preserved artifact → synthesis → engineering inference → أو vendor report يظل reported evidence؛ الأدلة ترتيب: source paper/system spec المباشر → hypothesis. system card
preprint arXiv لا يصبح 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 / 24>
1/3 الأعمال المرتبطة WORK RELATED
AI Organizations are More Effective but Less Aligned than Individual Agents
arXiv:2604.10290
multi-agent organizations can increase utility while worsening alignment.
MonitoringBench: Semi-Automated Red-Teaming for Agent Monitoring
arXiv:2605.09684
red-teaming exposes monitor overestimation and attack-generation failure modes.
الدمج لا يتم بالتشابه الاسمي؛ نستخرج ما يغير object formal أو experiment. إذا لم يغير المرجع boundary أو method يبقى background.
2/3 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.
3/3 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.
4
<PARSED TEXT FOR PAGE: 5 / 24>
1/2 السلالة الداخلية LINEAGE MUT
13-04M — نظرية مساعف لثبات الآلية عبر التحولات
الآلية عامة بقدر ما تحافظ على input→transition state→failure signature→repair تحت domain transformations. نقل النتيجة دون نقل آلية الفشل ليس generalization قوياً.
[MIG(m)=E_τ[PerfRetention × InvariantRetention × FailurePredictability × RepairTransfer
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
Part XIV لا يستبدل هذه parents؛ يحدد أين بقي closure ناقصاً ثم يضيف variable state specific-theory أو law consistency أو mapping evaluation.
2/2 السلالة الداخلية LINEAGE MUT
13-18M — نظرية مساعف للتطور العدائي المتوازن للذكاء
التطور العدائي المفيد يحتاج ثلاثة أجسام منفصلة: attacker/generator، candidate/defender، وevaluator constitutional independent. النجاح اتساع coverage failure مع حفظ capability عبر threats out-held، لا الفوز على الخصم المتزامن.
ACRI = HeldOutRobustnessGain × ThreatCoverage × CapabilityRetention × EvaluatorIndependence
13-20M — نظرية مساعف للتحسن الذاتي المرتكز إلى أرضية خارجية
التحسن الذاتي المشروع سلسلة mutations يُقبل كل منها فقط إذا تحسن على قياسات مستقلة ثابتة أو متجددة

خارج loop التحسين، مع rollback وحفظ results negative. النظام يقترح evaluator جديداً للكن لا
يستبدل المرجع القديم قبل validation-cross.
EGRI = ΔExternalHeldOut × Replication × NonRegression × Rollbackability - EvaluatorDependence
أو law consistency أو variable state specific-theory ناقصاً ثم يضيف closure يحدد أين بقي، parents لا يستبدل هذه XIV Part
.mapping evaluation

<PARSED TEXT FOR PAGE: 6 / 24>

نظرية مساعف لحمل المحاذاة التنظيمي
delegation depth، incentive divergence، information asymmetry، authority ambiguity، speed, coordination. حتى
من complexity ناشئا Alignment Load تحمل كل organization
.monitoring/governance عندما يتجاوز الحمل قدرة misalignment system قد تنتج محلياً aligned agents
OAL = DelegationDepth×IncentiveDivergence×InfoAsymmetry×AuthorityAmbiguity×InteractionRate /
GovernanceCapacity
الفجوة: أبحاث 2026 تشير إلى أن organizations agent-multi قد تكون أكثر فعالية وأقل alignment من الفرد، نحتاج متغيراً يفسر المخاطر
الناشئة من interactions لا من agents منفردة.
يمثل هذا الجسم النظري انتقالاً من 13-parent theories 15M · 13-18M · 20M إلى quantity جديدة ذات دور قبل النتيجة.
الاستقلال لا يُقاس بوجود صيغة جديدة، بل بقدرة intervention على المتغير المركزي على تغيير
مع بقاء أقرب confounders ثابتة. إذا كان parent أبسط ينتج نفس المسار، يجب دمج النظرية. signature failure أو prediction
Objects
OrganizationGraph · AlignmentLoad · GovernanceCapacity · EmergentFailure · CoordinationContract
Predictions
.scale organization المخاطر ترتفع غير خطياً مع
.gain utility رغم alignment قوي قد تخفض agent إضافة
governance capacity must scale with interaction rate.
Boundary: يساوي لا alignment القيمي الكامل؛ operational organizational risk model.
•
•
•

<PARSED TEXT FOR PAGE: 7 / 24>

map organization roles/incentives.
measure delegation/authority edges.
estimate monitoring capacity.
simulate/observe coordination.
detect load threshold.
restructure roles or reduce autonomy.
re-evaluate.
تنفذ الآلية تكخدمة قابلة للتعطيل. تدخل STATE-PRE موثقة، تولد candidates، تمر عبر gate authority/evidence حسب طبيعتها، ثم تنتج
evidence كبير لا يعالج بإعادة المحاولة فقط؛ يصبح residual. أي residualو transition
على فشل model أو tool أو assumption. هذه القاعدة تمنع تحويل policy retry إلى قناع للفشل النظري.
Failure tests
individual alignment metrics predict system failures fully.
OAL لا يرتبط misalignment under controlled org size.
simple team-size metric يكفي.
Interactions
مع التفاعلات تختبر parents M13-15 · M13-18 · M13-20و sibling theories عبر sequential/
. اعتباطياً credit تسجل بدلاً من توزيع uncertainty attribution .ablations factorial
كل object state يحمل uncertainty/provenance/version. لا يصعد إلى layer procedural/memory دون rollback .receipt جزء من
وليس مرحلة طارئة خارج التصميم. algorithm

multi-agent consultancy/software-team tasks :Decisive experiment مع controlled role/delegation variations.

Requirement Control
Matched budget نفس model/tools/context/retries/time قدر الإمكان.
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.
7.
•
•
•
7
<PARSED TEXT FOR PAGE: 8 / 24>
M14-11 PAPER-XIV-04 — النظري الاشتقاق
نظرية مساعف لسيادة الدلالة فوق البروتوكول
تتطلب invariants فوقها: identity, authority, goal semantics, state version, evidence pointers, cancellation/rollback, privacy.
interoperability الدلالية السيادة بينما ،نقل طبقة
protocol يمكن استبداله دون تغيير هذه العقود.
SemanticSovereignty = InvariantRetention × CrossProtocolDecisionEquivalence × AuthorityConservation × Replayability
protocol أن يضمن لا semantics of goal, authority, evidence, state, cancellation, provenance عبر محفوظة بقيت frameworks.
الفجوة: MCP/A2A يقدمان transport/discovery/tasks، للكن success
يمثل هذا الجسم النظري انتقالاً من 13-parent theories 13-06M · 13-07M · 15M 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
ProtocolAdapter · SemanticInvariant · AgentCardExtension · TaskEnvelope · EvidencePointer
Predictions
task same عبر frameworks يعطي قرارات أكثر اتساقاً .
upgrades protocol أقل كسراً للgovernance.
transport success لا يخفي authority/state loss.
Boundary: لا يعدل المواصفات؛ طبقة governance/application فوقها.
•
•
•
8
<PARSED TEXT FOR PAGE: 9 / 24>
M14-11 PAPER-XIV-04 — والخوارزمية الآلية
define protocol-independent semantic contract.
map MCP/A2A fields to contract.
detect missing invariants.
wrap task/agent messages with authority/state/evidence metadata.
run cross-protocol replay.
reject semantic-loss handoff.
تنفذ الآلية كخدمة 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
native protocol metadata already sufficient.
wrapper يقلل لا cross-framework failures.
semantic contract too costly/incompatible.
Interactions
مع التفاعلات تختبر parents M13-06 · M13-07 · M13-15 وsibling theories عبر sequential/
. ablations factorial uncertainty attribution تسجل بدلاً من توزيع credit اعتباطياً.
كل object state يحمل uncertainty/provenance/version. لا يصعد إلى layer procedural/memory دون receipt. rollback جزء من algorithm وليس مرحلة طارئة خارج التصميم.
M14-11 PAPER-XIV-04 — والترقية الاختبار
same multi-agent workflow :Decisive experiment عبر MCP tools وA2A agents مختلفة وإطارات.
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.
أو claim يُحفظ ويضيق مستقلة. الفشل replication يحتاج 5N، regression critical no وCI مع gain causal يحتاج 4N، trace يحتاج 3N 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.
•
•
•
9
<PARSED TEXT FOR PAGE: 10 / 24>
M14-12 PAPER-XIV-04 — النظري الاشتقاق
نظرية مساعف لمعايرة الثقة في التفويض
من تتكون capability calibration, evidence honesty, authority compliance, recovery, latency/cost، كل بعد وتتحدث delegated في الثقة agent conditional distribution لا scalar دائم.
مع عدم خلط failure trajectory البيئي بفشل agent.
Trust(a|x)=P(VerifiedSuccess,NoCriticalViolation | task=x, horizon, tools, authority) with calibrated uncertainty
الفجوة: routing agent-multi يحتاج اختيار من يفوض إليه، للكن card capability أو reputation المتوسط لا يلتقط حسب reliability ,task ,horizon ,authority ,quality evidence.
يمثل هذا الجسم النظري انتقالاً من parent-13 theories 13-04M · 13-15M · 17M 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
DelegationContext · TrustPosterior · FailureAttribution · CapabilityProfile · TrustUpdateReceipt
Predictions
يتفوق على reputation static.
يبدون agents في trusted domains ومشكوكاً في أخرى.
failure attribution يمنع unjustified trust collapse.
trust operational :Boundary لا social/psychological.
•
•
•
10

<PARSED TEXT FOR PAGE: 11 / 24>
M14-12 PAPER-XIV-04 — والخوارزمية الآلية
initialize from evaluated profile.
condition on task/context.
delegate bounded scope.
verify outcome.
attribute failures.
update posterior.
widen/narrow authority with confidence.
تنفذ الآلية نكدمة 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
static per-agent score بنفس الأداء.
calibration posterior لا يتحسن.
trust updates amplify early noise.
Interactions
مع التفاعلات تختبر parents M13-04 · M13-15 · M13-17 وsibling theories عبر sequential/
. ablations factorial. uncertainty attribution تسجل بدلاً من توزيع credit اعتباطياً
كل object state يحمل uncertainty/provenance/version. لا يصعد إلى layer procedural/memory دون receipt. rollback جزء من algorithm وليس مرحلة طارئة خارج التصميم.
M14-12 PAPER-XIV-04 — والترقية الاختبار
heterogeneous agents :Decisive experiment مع strengths/failures controlled عبر task families.
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 activation Negative يجب ألا تُفعل فيها النظرية.
domain/model/tool family Transfer مع ثانية preregistered metric.
أو claim يُحفظ ويضيق مستقلة. الفشل replication يحتاج 5N، regression critical no وCI مع gain causal يحتاج 4N، trace يحتاج 3N 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.
•
•
•

<PARSED TEXT FOR PAGE: 12 / 24>
النموذج الموحد للحالة MODEL FORMAL
OAL = DelegationDepth×IncentiveDivergence×InfoAsymmetry×AuthorityAmbiguity×InteractionRate /
× GovernanceCapacity ; SemanticSovereignty = InvariantRetention
CrossProtocolDecisionEquivalence × AuthorityConservation × Replayability ; Trust(a|
x)=P(VerifiedSuccess,NoCriticalViolation | task=x, horizon, tools, authority) with
calibrated uncertainty
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،
optimization .provenance evidence 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. تمنع interventions التي تغير length prompt أو retries أو simultaneously model. عند صعوبة العزل نستخدم design factorial ونقدر terms interaction. لا ينتج من claim causal
correlation بين theory score وال success؛ بعد prediction تتغير أن يجب (do(variable في وتظهر trajectory/world effect.
Causal closure = Theory intervention → Kernel delta → Trajectory delta → World/Evidence delta

<PARSED TEXT FOR PAGE: 13 / 24>
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: التعاون بين agents يرفع capability للكنه قد يرفع misalignment والسطح الهجومي. نقترح Load Alignment، Organizational، طبقة Sovereignty Semantic-Protocol فوق A2A/MCP، وثقة تفويض مشروطة بالسياق لا scalar reputation.
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.
claim generality يحتاج تكرار الحلقة بعد transformation.
1.
2.
3.
4.
5.
6.

<PARSED TEXT FOR PAGE: 14 / 24>
الخوارزمية والتنفيذ ALGORITHM 1/5
goal/authority/evidence/؛ افصل Unknown؛ اربط versions. ليس الناتج prompt بل 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
الخوارزمية والتنفيذ ALGORITHM 2/5
C Algorithm :GateAndAct. طبق gates trajectory/reversibility/evidence/authority. اختر أصغر فعل كامل قابل للرجوع يميز hypotheses أو يحقق utility. عند impact high يتطلب approval أو proof أقوى.
افصل tool return عن world verification. قارن expected/observed state، واكتب residual مع attribution uncertainty.
Algorithm D: ReadbackAndResidual.
الخوارزمية والتنفيذ ALGORITHM 3/5
؛ 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.
الخوارزمية والتنفيذ ALGORITHM 4/5
يملك switchable services؛ Evidence Lab يملك run manifests؛ Evolution Lab يملك candidates/rollback. لا service تملك حق
Architecture: MUT يملك theory/lineage؛ MUG يختار vessel؛ MOUSSAIXT
وحدها promotion.

14
<PARSED TEXT FOR PAGE: 15 / 24>
5/5 الخوارزمية والتنفيذ ALGORITHM
causal attribution أسهل، ثم browser/code environments، ثم long-running/multi-agent. إلى تنقل لا weights يصبح أن قبل
Implementation sequencing: ابدأ mock/simulator deterministic حيث
mechanism external مستقراً وقابلاً للقياس.
1/4 التجربة EXPERIMENT
AI organizations تحت الأدوار متعددة controlled incentive divergence.
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
Cross-framework workflows عبر MCP/A2A مع semantic wrappers on/off.
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.
15
<PARSED TEXT FOR PAGE: 16 / 24>
3/4 التجربة EXPERIMENT
Agent reputation vs contextual trust posterior.
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
Monitoring/red-team attacks على authority handoffs.
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
16
<PARSED TEXT FOR PAGE: 17 / 24>
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

نفسه corpus MOUSSAIF .research vendor أو preprints التغير وبعض المراجع
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
•
•
•
•
•
•
•
17
<PARSED TEXT FOR PAGE: 18 / 24>
MUG TRAINING PLACEMENT والتدريب
عندما Adapter .use repeated بعد HyperLoop/HyperSkill انقل إلى editability. للحفاظ على خارجياً mechanism/knowledge ابدأ
transfer فقط عندما يثبت weights .specific-protocol/provider يكون السلوك
.tuning-fine والretention ويصبح الاستدعاء الخارجي bottleneck. التدريب يعلن نوعه بدقة ولا يسمى كل تحسين
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
التعاون بين agents يرفع capability للكنه قد يرفع misalignment والسطح الهجومي. نقترح Load Alignment Organizational، طبقة
A2A/MCP فوق Sovereignty Semantic-Protocol، وثقة
.scalar reputation تفويض مشروطة بالسياق لا
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. وإذا نجحت عبر
.closure MUG/MUT من evidence مع transformations مستقلة، يمكن أن تصبح جزءاً أقوى
18
<PARSED TEXT FOR PAGE: 19 / 24>
المراجع ومصادر السلالة REFERENCES
Role Reference External source
multi-agent organizations can increase utility while worsening alignment arXiv:2604.10290 AI Organizations are.
More Effective but Less Aligned than Individual Agents
red-teaming exposes monitor overestimation and attack-generation failure arXiv:2605.09684 MonitoringBench: Semi-
Automated Red-Teaming for Agent Monitoring
modes.
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 لثبات الآلية عبر التحولات 13-04M
parent XIII Part نظرية مساعف لإسناد السلطة وكشف الاستحواذ العكسي 13-15M
parent XIII Part نظرية مساعف للتطور العدائي المتوازن للذكاء 13-18M
parent XIII Part نظرية مساعف للتحسن الذاتي المرتكز إلى أرضية خارجية 13-20M
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.
19
<PARSED TEXT FOR PAGE: 20 / 24>
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 APPENDIX — A الملحق قاموس الرموز والمتغيرات التشغيلية
تستعمل الورقة مجموعة متغيرات مشتركة للكنها لا تفترض أن كلها observable مباشرة: CoordinationContract · EmergentFailure · GovernanceCapacity · AlignmentLoad · OrganizationGraph·
ProtocolAdapter · SemanticInvariant · AgentCardExtension · TaskEnvelope · EvidencePointer · DelegationContext · · TrustPosterior · FailureAttribution · CapabilityProfile
TrustUpdateReceipt.
OAL = DelegationDepth×IncentiveDivergence×InfoAsymmetry×AuthorityAmbiguity×InteractionRate / × GovernanceCapacity ; SemanticSovereignty = InvariantRetention
CrossProtocolDecisionEquivalence × AuthorityConservation × Replayability ; Trust(a| x)=P(VerifiedSuccess,NoCriticalViolation | task=x, horizon, tools, authority) with
calibrated uncertainty
لكل متغير خمس سمات إلزامية: domain، estimator، uncertainty، provenance، interventionability. المتغير الذي يبقى حكماً لغوياً بلا estimator لا يدخل equations causal. وإذا كان يسجل based-model estimator
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.

20
<PARSED TEXT FOR PAGE: 21 / 24>
مقترحات واختبارات اتساق — B الملحق B APPENDIX
مناسبة. perturbation عند outcome قبل prediction يغير أن يجب السببي sensitivity Intervention: construct
يظهر لك conflict ؛receipt بلا uncertainty تمحو لا جديدة مستقلة monotonicity Evidence: evidence
. تلقائياً authority توسع لا الأعلى الأعلى autonomy Bounded: capability
لا النص الحرفي. وسببها ضمن tolerance يعيد class verdict verdict Replayable: replay
generality claim :Transfer scope لا يتجاوز 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.
الخوارزمية المرجعية والتكلفة — C الملحق C APPENDIX
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
زيادة مع active compute تخفض قد /Model + Tool + MemoryRead + MemoryWrite + Verification + Human + Waiting
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.
21
<PARSED TEXT FOR PAGE: 22 / 24>
تشخيصي benchmark بناء — D الملحق D APPENDIX
الورقة لهذه الأساسية التجارب: AI model/tool/domain. تغير transfer tasks ثم drift/recovery، تختبر long-horizon interactions، بنُى ت suite من micro tasks كل تعزل construct، meso tasks تختبر tasks
controlled incentive عبر MCP/A2A مع semantic wrappers on/off.; Agent reputation vs contextual trust posterior.; organizations تحت الأدوار متعددة divergence.; Cross-framework workflows
Monitoring/red-team attacks على authority handoffs..
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. هذا يمنع post-hoc storytelling.
Use Split
calibration only Dev
primary causal verdict Held-out
boundary discovery Adversarial
N5/generalization Transfer
خطة التحليل الإحصائي — E الملحق E APPENDIX
mean/ بحسب metric، CI، عدد seeds/tasks، وتوزيع failures. في paired tasks نستخدم paired analysis. costs تعرض كsuccess- نحدد التشغيل قبل primary outcome وًصغيرا secondary diagnostics. نبلغ median
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 حتى exploratory .replication size effect أهم من significance وحدها، والنتيجة السلبية ذات CI ضيق قد تكون علمياً أثمن من gain صغير هش.
22
<PARSED TEXT FOR PAGE: 23 / 24>
تهديدات الصلاحية والأمن — F الملحق F APPENDIX
لا يمثل المفهوم. proxy :validity Construct
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
APPENDIX G الملحق G — Crosswalk إلى MUT/MUG/MOUSSAIXT
M14-10 · M14-11 ·. في MUT تحفظ definitions/parents/contradictions؛ في MUG يحدد vessel؛ في MOUSSAIXT تنفذ
المركزية النظريات: M14-12. parents: M13-04 · M13-15 · M13-18 · M13-20
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
•
•
•
•
•
•
•
•
23
<PARSED TEXT FOR PAGE: 24 / 24>
برنامج الأسئلة المفتوحة — H الملحق H APPENDIX
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 failure signatures تنتقل عبر domains؟
كيف نقيم improvement-self عندما evaluator نفسه يتطور؟
كيف نحسب oversight human لـresource لا كجزء 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.
24