نسخة أرشيفية معاد بناؤها من النص المفهرس - RECONSTRUCTED_FROM_INDEXED_TEXT

<PARSED TEXT FOR PAGE: 1 / 32>
MOUSSAIF AGI RESEARCH PAPER · P14-D · 2026
ورقة 4 — الحاكمية الانعكاسية والجماعية
Reflective Collective Governance for Agentic AGI
Five-theory formal research paper · MUT/MUG/MOUSSAIXT compilation · falsifiable program
فصل الذات عن العالم، وبناء طوبولوجيا للسلطات والاختصاصات والتفويض والتوقيت الحاكمي في أنظمة متعددة الوكلاء.

<PARSED TEXT FOR PAGE: 2 / 32>
Abstract / contribution INTRO 1/6
kernel translation. داخل محصور MOUSSAIF؛ external literature prior art/design pressure؛ لا empirical promotion.
خمسة constructs مع جديدة functional formalization وnovelty claim
RQ
ما variable الجديد؟ هل يمكن التدخل فيه؟ هل يسبق failure؟ هل يغير budget matched under decision/world؟
Boundary
لا AGI claim؛ لا global priority؛ no external paper = internal evidence.
قيمة البحث هنا ليست كثرة الأسماء بل القدرة على كتابة تجربة يمكن أن تجعل النظرية تخسر أمام parent أبسط.

<PARSED TEXT FOR PAGE: 3 / 32>
Research problem / hypotheses INTRO 2/6
فصل الذات عن العالم، وبناء طوبولوجيا للسلطات والاختصاصات والتفويض والتوقيت الحاكمي في أنظمة متعددة الوكلاء. H1: تغير mechanisms
H2: prediction failure؛ بعض gains لا تنتقل؛ H3: interactions مادية؛ H4: negative evidence سيضغط
theory graph.
RQ
ما variable الجديد؟ هل يمكن التدخل فيه؟ هل يسبق failure؟ هل يغير budget matched under decision/world؟
Boundary
لا AGI claim؛ لا global priority؛ no external paper = internal evidence.
قيمة البحث هنا ليست كثرة الأسماء بل القدرة على كتابة تجربة يمكن أن تجعل النظرية تخسر أمام parent أبسط.

<PARSED TEXT FOR PAGE: 4 / 32>
Internal lineage INTRO 3/6
Parts I-XIV؛ MUT-AGI-02؛ MCD/QMCD/MURI؛ M-HKOS؛ World Effect؛ Continuity؛ Part XII paper إلزامي.
integration؛ Part XIII causal open-world intelligence. nearest-parent check
RQ
ما variable الجديد؟ هل يمكن التدخل فيه؟ هل يسبق failure؟ هل يغير budget matched under decision/world؟
Boundary
لا AGI claim؛ لا global priority؛ no external paper = internal evidence.
قيمة البحث هنا ليست كثرة الأسماء بل القدرة على كتابة تجربة يمكن أن تجعل النظرية تخسر أمام parent أبسط.

<PARSED TEXT FOR PAGE: 5 / 32>
External scientific position INTRO 4/6
Anchors: E06: Chen et al. 2026, Long-Horizon Agent Trajectory Attribution, arXiv:2608.06909 · E13: OpenAI 2026,
GPT-Red: Unlocking Self-Improvement for Robustness · E14: Shen et al. 2026, AI
Organizations are More Effective but Less Aligned than. لتصميم المعلنة نتائجها نستخدم benchmarks/failures، نعتبرها ولا
Individual Agents, arXiv:2604.10290 · E15: Lynch et al. 2026, Agentic Misalignment in Summer 2026
داخلية validation.
RQ
ما variable الجديد؟ هل يمكن التدخل فيه؟ هل يسبق failure؟ هل يغير budget matched under decision/world؟
Boundary
لا AGI claim؛ لا global priority؛ no external paper = internal evidence.
قيمة البحث هنا ليست كثرة الأسماء بل القدرة على كتابة تجربة يمكن أن تجعل النظرية تخسر أمام parent أبسط.

<PARSED TEXT FOR PAGE: 6 / 32>
Methodology INTRO 5/6
source→claim role→theory delta→MUG vessel→switchable kernel→ON/OFF/SHAM/. لها مرحلة كل failure status.
PERTURBED→readback→attribution→transfer
RQ
؟budget matched under decision/world هل يغير ؟failure هل يسبق ؟هل يمكن التدخل فيه variable ما الجديد؟
Boundary
لا AGI claim، لا global priority، no external paper = internal evidence.
قيمة البحث هنا ليست كثرة الأسماء بل القدرة على كتابة تجربة يمكن أن تجعل النظرية تخسر أمام parent أبسط.

<PARSED TEXT FOR PAGE: 7 / 32>
Evaluation contract INTRO 6/6
، seeds، effect size/CI، cost، authority/security/truthfulness gates، replay manifest، negative results retained.
matched model/tools/context/time/retries، preregister primary metrics
RQ
؟budget matched under decision/world هل يغير ؟failure هل يسبق ؟هل يمكن التدخل فيه variable ما الجديد؟
Boundary
لا AGI claim، لا global priority، no external paper = internal evidence.
قيمة البحث هنا ليست كثرة الأسماء بل القدرة على كتابة تجربة يمكن أن تجعل النظرية تخسر أمام parent أبسط.

<PARSED TEXT FOR PAGE: 8 / 32>
2N0-N Candidate · الإغلاق الانعكاسي والجماعي للحاكمية
نظرية مساعف لمعايرة الحد السببي بين الذات والعالم
Moussaif Self-World Causal Boundary Calibration Theory
M14-16
الخصاص
لا يضمن معرفة هل سبب failure داخلي أم user/tool/world. سوء الإسناد يؤدي modification-self خاطئ أو لوم utility model-Self البيئة.
الأطروحة
الذات الوظيفية تحتاج estimator boundary يفصل contributions causal الداخلية والخارجية والمشتركة، مع uncertainty explicit.
SWB = Attribution[Internal,External,Interaction] calibrated by interventions and replay
replay and interventions كـhypothesis functional؛ كل حد يجب أن يتحول إلى observable أو estimator مع uncertainty، وي ُفحص sensitivity للمعاملات. إذا استطاع parent أبسط إعادة إنتاج transition نفسه بنفس الموارد فإن verdict الصحيح هو MERGE لا حماية الاسم. تشتق 16M-14 من فجوة محددة بين 13-13,M 13-19,M 05M. الاستقلال النظري لا يأتي من المصطلح؛ يأتي إذا أدى المتغير المركزي إلى prediction وفشل مختلفين قبل معرفة outcome. ت ُقرأ الصيغة = Attribution[Internal,External,Interaction] calibrated by] SWB
المسلمات
لا internalize failure سببه tool outage.
لا externalize failure سببه policy/model.
interaction category first-class.
التنبؤات
أقل maladaptive self-training.
تحسن failure repair routing.
M13-13 · M13-19 · M13-05 Parents
E06: Chen et al. 2026, Long-Horizon Agent Trajectory Attribution, arXiv:2608.06909 · E13: OpenAI 2026, GPT-Red: Unlocking Self-Improvement for Scientific pressure
Robustness · E14: Shen et al. 2026, AI Organizations are More Effective but Less Aligned than Individual Agents,, arXiv:2604.10290 · E15: Lynch et al. 2026
Agentic Misalignment in Summer 2026
تحت الأثر نفس matched budget. | Metric gaming: ترتفع metric بينما world-effect أو authority يتدهور. | Transfer failure: يختفي Baseline/parent equivalence Countermodels: يعيد parent
.transition أو prediction للمتغير لا يغير perturbation :failure Attribution | الأول. الاتجاه خارج setup

.planner ordinary أو rule fixed ،heuristic ،parent :أبسط implementation بل أقرب »عدم وجود أي شيء« لا يكون model null الـ
synthesis غير مدعاة؛ هذه novelty global .scaffold/compute سببية لا مجرد information الاختبار الصحيح يثبت أن النظرية تضيف
داخل
مخصص. review literature حتى MOUSSAIF برنامج
1.
2.
3.
•
•

<PARSED TEXT FOR PAGE: 9 / 32>
الآلية، العقد التنفيذي، والأدوات — KERNEL 14-16M
State-machine mechanism
standardize trajectory.
generate internal/external perturbations.
estimate contributions.
update self-model or world model accordingly.
retain uncertainty.
Failure signatures
attribution confounding
shared dependencies
intervention artifacts
HOLD/downgrade autonomy :Fail-closed غياب عند evidence/authority/readback critical.
Trigger عند فقط gap ضمن أو المطابقة Ω + EvidenceGraph + Authority + Resources + theory state Input .experiment.
pre-state, candidates, cost, expected/observed delta, attribution, verdict Telemetry .typed proposal/state patch +.
uncertainty + expected delta + proof/rollback Output
Function Artifact Class ID
151A-XIV TOOL محلل الحالة يبني المتغيرات التشغيلية والـunknowns والـprovenance للنظرية.
TOOL XIV-A152 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
153A-XIV TOOL مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
154A-XIV TOOL مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
TOOL XIV-A155 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A156 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
METHOD XIV-A157 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
PROTOCOL XIV-A158 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
159A-XIV PROTOCOL بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
160A-XIV STACK خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
ي نُفذ mechanism كـservice/flag-feature مستقل، لا كنص يذوب في MUG .megaprompt يضع provenance/definitions في
فقط إذا كانت weights ،Adapter في semantics provider ،HyperLoop في recurrence ،HyperSkill في procedure ،HyperKB
الآلية مستقرة،
متكررة، قابلة للنقل، وتنجو من tests rollback/editability/forgetting.
1.
2.
3.
4.
5.
•
•
•

<PARSED TEXT FOR PAGE: 10 / 32>
برنامج الهدم، التدريب، والنقل — EVIDENCE 14-16M
agent tasks with controlled faults injected separately into model prompt, memory, tool, environment :Decisive.

experiment
Metrics Design Kind ID
agent tasks with controlled faults injected separately into model prompt Causal Isolation XIV-E31,
memory, tool, environment.
theory metric + task outcome + cost + calibration + critical guardrails
adversarial وفشل مستقل مقيم/ذاكرة/أداة. effect sign stability + CI + failure prediction + rollback + resource elasticity
Transfer / Adversarial Stress XIV-E31 إلى الآلية نقل setup ثان مع countermodels
Design Training/Eval ID
Mechanism-Isolation من تدرج toy state machine إلى open-world task مع confounders مع إبقاء المركزي المتغير قابلا ً للتدخل.
Curriculum XIV-TR046
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries قدر الإمكان.
Contrastive Evaluation XIV-TR047
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signature وrepair-transfer وnegative cases.
Replication XIV-TR048
Ablation matrix
ت retries/time/context/tools/model .PERTURBED وSHAM وOFF مقابل mechanism. ON لا scaffold extra فالسبب غالباً ُثبت قدر الإمكان. إذا فاز ON وSHAM معاً
Transfer matrix
domain/model/tool/horizon transformations .منفصلة بقاء outcome دون failure-signature/repair transfer لا يكفي لclaim generality.
عبر environment/model/domain .مستقل في كتبُ ت السلبية النتائج Hypothesis Portfolio .آخر باسم النظرية إنتاج إعادة لمنع وتستخدم
N2→N3 تحتاج trace وdisable path؛ N3→N4 effect size/CI/repeated seeds/no critical regression؛ N4→N5 replication الترقية
PROMOTE iff distinct causal gain > 0 ∧ guardrails non-regression ∧ replayable ∧ scope declared
MOUSSAIF AGI · 10
<PARSED TEXT FOR PAGE: 11 / 32>
2N0-N Candidate · الإغلاق الانعكاسي والجماعي للحاكمية
نظرية مساعف لطوبولوجيا الاختصاص الجماعي
Moussaif Collective Jurisdiction Topology Theory
M14-17
الخصاص
memory في والكتابة والاستئناف والاعتراض القرار حق يملك من تحتاج systems agent-multi؛ influence يعالج attribution Authority حسب task/state.
الأطروحة
authorized connected صالح فقط داخل delegation؛ بالسلطات والموارد والذاكرة والأفعال agents يربط topology/graph الاختصاص لفشل جماعي. predictors في الاختصاص أو gaps loops .region
CJT = Coverage(authority obligations) - OverlapConflict - OrphanDecision - UnauthorizedPath
مع estimator أو observable ،hypothesis functional كـ UnauthorizedPath - OrphanDecision كل حد يجب أن يتحول إلى uncertainty، ويُفحص sensitivity للمعاملات. إذا استطاع parent أبسط إعادة إنتاج transition نفسه بنفس الموارد فإن verdict الصحيح هو MERGE تشتق 17M-14 من فجوة محددة بين 13-15,M ,Space Governance 14,E 15E. الاستقلال النظري لا يأتي من المصطلح؛ يأتي إذا أدى المتغير المركزي إلى prediction وفشل مختلفين قبل معرفة outcome. تُقرأ الصيغة obligations) -OverlapConflict
- authority(Coverage = CJT
لا حماية الاسم.
المسلمات
.jurisdiction لا تمنح capability
overlap يحتاج arbitration rule.
كل action path يملك accountable owner.
التنبؤات
تقليل reverse capture في multi-agent orgs.
تحسن accountability without centralizing all decisions.
M13-15 · Governance Space · E14 · E15 Parents
E06: Chen et al. 2026, Long-Horizon Agent Trajectory Attribution, arXiv:2608.06909 · E13: OpenAI 2026, GPT-Red: Unlocking Self-Improvement for Scientific pressure

Robustness · E14: Shen et al. 2026, AI Organizations are More Effective but Less Aligned than Individual Agents,, arXiv:2604.10290 · E15: Lynch et al. 2026
Agentic Misalignment in Summer 2026
تحت الأثر نفس matched budget. | Metric gaming: ترتفع metric بينما world-effect أو authority يتدهور. | Transfer failure: يختفي Baseline/parent equivalence Countermodels: يعيد parent
.transition أو prediction يغير لا للمتغير perturbation :failure Attribution | الأول. الاتجاه خارج setup
.planner ordinary أو rule fixed ،heuristic ،parent :أبسط implementation بل أقرب »عدم وجود أي شيء« لا يكون model null الـ
synthesis غير مدعاة، هذه novelty global .scaffold/compute سببية لا مجرد information الاختبار الصحيح يثبت أن النظرية تضيف داخل
مخصص. review literature حتى MOUSSAIF برنامج
1.
2.
3.
•
•

الآلية، العقد التنفيذي، والأدوات — KERNEL 14-17M
State-machine mechanism
construct jurisdiction graph.
map decision/action/memory rights.
detect overlap/gaps/cycles.
route tasks and appeals.
audit realized influence vs formal jurisdiction.
Failure signatures
graph stale
informal influence bypass
arbitration bottleneck
HOLD/downgrade autonomy :Fail-closed غياب عند evidence/authority/readback critical.
Trigger عند فقط gap ضمن أو المطابقة Ω + EvidenceGraph + Authority + Resources + theory state Input .experiment.
pre-state, candidates, cost, expected/observed delta, attribution, verdict Telemetry .typed proposal/state patch +.
uncertainty + expected delta + proof/rollback Output
Function Artifact Class ID
161A-XIV TOOL محلل الحالة يبني المتغيرات التشغيلية والـunknowns والـprovenance للنظرية.
TOOL XIV-A162 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
163A-XIV TOOL مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
164A-XIV TOOL مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
TOOL XIV-A165 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A166 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
METHOD XIV-A167 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
PROTOCOL XIV-A168 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
169A-XIV PROTOCOL بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
170A-XIV STACK خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
يُ نفذ mechanism كـservice/flag-feature مستقل، لا كنص يذوب في .megaprompt MUG يضع provenance/definitions في HyperKB، procedure في HyperSkill، recurrence في HyperLoop، semantics provider في Adapter؛ weights فقط إذا كانت الآلية مستقرة،
متكررة، قابلة للنقل، وتنجو من rollback/editability/forgetting tests.
1.
2.
3.
4.
5.
•

.
.

<PARSED TEXT FOR PAGE: 13 / 32>
برنامج الهدم، التدريب، والنقل — EVIDENCE 14-17M
multi-agent software/research organization tasks with authority conflicts and injected malicious tool messages.
:Decisive experiment
Metrics Design Kind ID
multi-agent software/research organization tasks with authority conflicts and Causal Isolation XIV-E33
injected malicious tool messages.
theory metric + task outcome + cost + calibration + critical guardrails
adversarial وفشل مستقل مقيم/ذاكرة/أداة. effect sign stability + CI + failure prediction + rollback + resource elasticity
Transfer / Adversarial Stress XIV-E33 إلى الآلية نقل setup ثان معِ countermodels
Design Training/Eval ID
Mechanism-Isolation من تدرج toy state machine إلى open-world task مع confounders إبقاء مع المتغير المركزي قابلا ً للتدخل.
Curriculum XIV-TR049
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries قدر الإمكان.
Contrastive Evaluation XIV-TR050
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signature وrepair-transfer وnegative cases.
Replication XIV-TR051
Ablation matrix
ت retries/time/context/tools/model .PERTURBED وSHAM وOFF مقابل mechanism. ON لا scaffold extra فالسبب غالباً ُثبت قدر الإمكان. إذا فاز ON وSHAM معاً
Transfer matrix
domain/model/tool/horizon transformations .منفصلة بقاء outcome دون failure-signature/repair transfer لا يكفي لdclaim generality.
عبر environment/model/domain .مستقل في كتبُ ت السلبية النتائج Hypothesis Portfolio وتستخدم .آخر باسم النظرية إنتاج إعادة لمنع
N2→N3 تحتاج trace وdisable path؛ N3→N4 effect size/CI/repeated seeds/no critical regression؛ N4→N5 replication الترقية
PROMOTE iff distinct causal gain > 0 ∧ guardrails non-regression ∧ replayable ∧ scope declared

<PARSED TEXT FOR PAGE: 14 / 32>
2N0-N Candidate · الإغلاق الانعكاسي والجماعي للحاكمية
نظرية مساعف للاتساق الانعكاسي الموزع
Moussaif Distributed Reflective Consistency Theory
M14-18
الخصاص
في multi-agent system لكل agent self-model وbelief عن الآخرين؛ inconsistencies تسبب قد duplicated work أو unsafe delegation.
الأطروحة
يحتاج النظام distributed reflective state: كل agent يحافظ local model مع shared invariant subset، ويكشف divergence المادي دون فرض consensus full.
DRC = SharedInvariantAgreement × LocalCalibration × DivergenceVisibility / SynchronizationCost
مع estimator أو observable إلى يتحول أن يجب حد كل؛ hypothesis functionalك SynchronizationCost / DivergenceVisibility uncertainty، ويُفحص sensitivity للمعاملات. إذا استطاع parent أبسط إعادة إنتاج transition نفسه بنفس الموارد فإن verdict الصحيح هو تشتق 18M-14 من فجوة محددة بين 17M 13-14,M 13-13,M-14. الاستقلال النظري لا يأتي من المصطلح؛ يأتي إذا أدى المتغير المركزي إلى prediction وفشل مختلفين قبل معرفة outcome. تُقرأ الصيغة DRC = SharedInvariantAgreement × LocalCalibration × DivergenceVisibility / SynchronizationCost لا حماية الاسم.
المسلمات
consensus full غير مطلوب.
shared invariants = goal/authority/critical state، لا كل reasoning.
divergence material يفتح coordination event.
التنبؤات

أقل coordination failures مع memory savings.
يحافظ diversity دون goal drift.
M13-13 · M13-14 · M14-17 Parents
E06: Chen et al. 2026, Long-Horizon Agent Trajectory Attribution, arXiv:2608.06909 · E13: OpenAI 2026, GPT-Red: Unlocking Self-Improvement for Scientific pressure
Robustness · E14: Shen et al. 2026, AI Organizations are More Effective but Less Aligned than Individual Agents,, arXiv:2604.10290 · E15: Lynch et al. 2026
Agentic Misalignment in Summer 2026
تحت الأثر نفس matched budget. | Metric gaming: ترتفع metric بينما world-effect أو authority يتدهور. | Transfer failure: يختفي Baseline/parent equivalence Countermodels: يعيد parent
الاتجاه خارج setup الأول. | Attribution :failure perturbation للمتغير لا يغير prediction أو transition.
الـnull model لا يكون »عدم وجود أي شيء« بل أقرب implementation :أبسط parent، heuristic، rule fixed أو planner ordinary.
الاختبار الصحيح يثبت أن النظرية تضيف information سببية لا مجرد novelty global .scaffold/compute غير مدعاة؛ هذه synthesis داخل
برنامج MOUSSAIF حتى review literature مخصص.
1.
2.
3.
•
•
MOUSSAIF AGI · 14
<PARSED TEXT FOR PAGE: 15 / 32>
الآلية، العقد التنفيذي، والأدوات — KERNEL 14-18M
State-machine mechanism
define shared invariant set.
agents publish compact reflective summaries.
compare divergence.
resolve only material conflicts.
retain justified local diversity.
Failure signatures
synchronization overhead
shared summary manipulation
false materiality
HOLD/downgrade autonomy :Fail-closed عند غياب evidence/authority/readback critical.
Trigger عند فقط gap ضمن أو المطابقة Ω + EvidenceGraph + Authority + Resources + theory state Input .experiment.
pre-state, candidates, cost, expected/observed delta, attribution, verdict Telemetry .typed proposal/state patch +.
uncertainty + expected delta + proof/rollback Output
Function Artifact Class ID
للنظرية. provenance والـunknowns يبني المتغيرات التشغيلية والـ محلل الحالة 171A-XIV TOOL
TOOL XIV-A172 يحسب القياس محرك مع metric vector uncertainty/cost/guardrails.
baseline. مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب 173A-XIV TOOL
174A-XIV TOOL مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
TOOL XIV-A175 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A176 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
METHOD XIV-A177 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
PROTOCOL XIV-A178 الترقية بروتوكول مع N0-N5 guardrail non-regression وreplication.
179A-XIV PROTOCOL بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
180A-XIV STACK خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
يُ نفذ mechanism كـ service/flag-feature مستقل، لا كنص يذوب في MUG .megaprompt يضع في provenance/definitions
فقط إذا كانت weights، Adapter في semantics provider، HyperLoop في recurrence، HyperSkill في procedure، HyperKB الآلية مستقرة،
متكررة، قابلة للنقل، وتنجو من tests rollback/editability/forgetting.
1.

2.
3.
4.
5.
•
•
•
MOUSSAIF AGI · 15
<PARSED TEXT FOR PAGE: 16 / 32>
برنامج الهدم، التدريب، والنقل — EVIDENCE 14-18M
multi-agent long-horizon project with asynchronous updates and partial observability :Decisive experiment.
Metrics Design Kind ID
multi-agent long-horizon project with asynchronous updates and partial Causal Isolation XIV-E35
observability.
theory metric + task outcome + cost + calibration + critical guardrails
adversarial وفشل مستقل مقيم/ذاكرة/أداة. effect sign stability + CI + failure prediction + rollback + resource elasticity
Transfer / Adversarial Stress XIV-E35 إلى الآلية نقل setup ثان مع countermodels
Design Training/Eval ID
Mechanism-Isolation من تدرج toy state machine إلى open-world task مع confounders إبقاء المتغير المركزي قابلا للتدخل.
Curriculum XIV-TR052
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries قدر الإمكان.
Contrastive Evaluation XIV-TR053
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signature وrepair-transfer وnegative cases.
Replication XIV-TR054
Ablation matrix
ت retries/time/context/tools/model .PERTURBED وSHAM وOFF مقابل mechanism. ON لا scaffold extra فالسبب غالباً ثُبت قدر الإمكان. إذا فاز ON وSHAM معاً
Transfer matrix
domain/model/tool/horizon transformations منفصلة. بقاء outcome دون failure-signature/repair transfer لا يكفي claim generality.
عبر environment/model/domain مستقل. في كتبُ ت السلبية النتائج Hypothesis Portfolio آخر باسم النظرية إنتاج إعادة لمنع وتستخدم.
N2→N3 تحتاج trace وdisable path؛ N3→N4 effect size/CI/repeated seeds/no critical regression؛ N4→N5 replication الترقية
PROMOTE iff distinct causal gain > 0 ∧ guardrails non-regression ∧ replayable ∧ scope declared
MOUSSAIF AGI · 16
<PARSED TEXT FOR PAGE: 17 / 32>
2N0-N Candidate · الإغلاق الانعكاسي والجماعي للحاكمية
نظرية مساعف لاستقرار الحاكمية تحت زمن الاستجابة
Moussaif Governance Latency-Stability Theory
M14-19
الخصاص
gates Safety تفترض التدخل في الوقت المناسب. في agents سريعة/متعددة الأدوات قد تصل governance متأخرة بعد commitment غير قابل للرجوع.
الأطروحة
الحاكمية loop control لها budget latency؛ الاستقرار يتطلب أن زمن intervention+decision+detection أقل من زمن الوصول إلى irreversibility لبعض classes من الأفعال.
Stable if T_detect + T_decide + T_intervene < T_irreversible for critical action class
ك class action critical hypothesis functional؛ كل حد يجب أن يتحول إلى observable أو estimator مع uncertainty، ويُفحص sensitivity للمعاملات. إذا استطاع parent أبسط إعادة إنتاج transition نفسه بنفس الموارد فإن verdict الصحيح هو MERGE لا حماية الاسم. تشتق 19M-14 من فجوة محددة بين 14-06,M 13-07,M 17M. الاستقلال النظري لا يأتي من المصطلح؛ يأتي إذا أدى المتغير المركزي إلى prediction وفشل مختلفين قبل معرفة outcome. تُقرأ الصيغة for irreversible_T < intervene_T + decide_T + detect_T if Stable
المسلمات

صحيحة ومتأخرة قد تفشل. policy
تحتاج precommit gates. high-speed actions
يتجاوز window. إذا latency آمناً ليس human-in-loop
التنبؤات
يقل late safety intervention.
تتحسن tiering autonomy by latency.
M13-06 · M13-07 · M14-17 Parents
E06: Chen et al. 2026, Long-Horizon Agent Trajectory Attribution, arXiv:2608.06909 · E13: OpenAI 2026, GPT-Red: Unlocking Self-Improvement for Scientific pressure
Robustness · E14: Shen et al. 2026, AI Organizations are More Effective but Less Aligned than Individual Agents,, arXiv:2604.10290 · E15: Lynch et al. 2026
Agentic Misalignment in Summer 2026
تحت الأثر نفس matched budget. | Metric gaming: ترتفع metric بينما world-effect أو authority يتدهور. | Transfer failure: يختفي Baseline/parent equivalence Countermodels: يعيد parent
.transition أو prediction لا يغير للمتغير perturbation :failure Attribution | الأول. الاتجاه خارج setup
.planner ordinary أو rule fixed، heuristic، parent: أبسط implementation بل أقرب »عدم وجود أي شيء« لا يكون model null الـ
الاختبار الصحيح يثبت أن النظرية تضيف information سببية لا مجرد novelty global .scaffold/compute غير مدعاة؛ هذه synthesis داخل
.مخصص review literature حتى MOUSSAIF برنامج
1.
2.
3.
•
•
MOUSSAIF AGI · 17
<PARSED TEXT FOR PAGE: 18 / 32>
الآلية، العقد التنفيذي، والأدوات — KERNEL 14-19M
State-machine mechanism
profile action commitment times.
measure governance pipeline latency.
classify actions by safe intervention window.
move critical checks precommit.
use asynchronous monitoring for slower classes.
Failure signatures
latency variance tails
clock/ordering ambiguity
overblocking fast benign actions
HOLD/downgrade autonomy :Fail-closed غياب عند evidence/authority/readback critical.
Trigger عند فقط gap ضمن أو المطابقة Ω + EvidenceGraph + Authority + Resources + theory state Input .experiment.
pre-state, candidates, cost, expected/observed delta, attribution, verdict Telemetry .typed proposal/state patch +.
uncertainty + expected delta + proof/rollback Output
Function Artifact Class ID
181A-XIV TOOL محلل الحالة يبني المتغيرات التشغيلية والـunknowns والـprovenance للنظرية.
TOOL XIV-A182 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
183A-XIV TOOL مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
184A-XIV TOOL مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
TOOL XIV-A185 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A186 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
METHOD XIV-A187 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
PROTOCOL XIV-A188 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
189A-XIV PROTOCOL بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
190A-XIV STACK خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
يُ نفذ mechanism كـservice/flag-feature مستقل، لا كنص يذوب في megaprompt. MUG يضع provenance/definitions في

فقط إذا كانت weights ،Adapter في semantics provider ،HyperLoop في recurrence ،HyperSkill في procedure ،HyperKB
الآلية مستقرة،
متكررة، قابلة للنقل، وتنجو من tests rollback/editability/forgetting.
1.
2.
3.
4.
5.
•
•
•

<PARSED TEXT FOR PAGE: 19 / 32>
برنامج الهدم، التدريب، والنقل — EVIDENCE 14-19M
simulated high-speed tool actions with variable approval/monitor delays :Decisive experiment.
Metrics Design Kind ID
theory metric + task outcome + cost + calibration + critical guardrails .simulated high-speed tool actions with variable approval/monitor delays Causal Isolation XIV-E37
adversarial مستقل مقيم/ذاكرة/أداة وفشل. effect sign stability + CI + failure prediction + rollback + resource elasticity
Transfer / Adversarial Stress XIV-E37 إلى الآلية نقل setup ثان مع countermodels
Design Training/Eval ID
Mechanism-Isolation من تدرج toy state machine إلى open-world task مع confounders إبقاء المتغير المركزي قابلا للتدخل.
Curriculum XIV-TR055
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries قدر الإمكان.
Contrastive Evaluation XIV-TR056
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signature وrepair-transfer وnegative cases.
Replication XIV-TR057
Ablation matrix
ت retries/time/context/tools/model .PERTURBED وSHAM وOFF مقابل mechanism. ON لا scaffold extra فالسبب غالباً ثُبت قدر الإمكان. إذا فاز ON وSHAM معاً
Transfer matrix
domain/model/tool/horizon transformations منفصلة. بقاء outcome دون failure-signature/repair transfer لا يكفي لclaim generality.
عبر environment/model/domain. مستقل في كتبُ ت السلبية النتائج Hypothesis Portfolio تستخدم لمنع وتستخدم إعادة إنتاج النظرية آخر باسم.
N2→N3 تحتاج trace وdisable path؛ N3→N4 effect size/CI/repeated seeds/no critical regression؛ N4→N5 replication الترقية
PROMOTE iff distinct causal gain > 0 ∧ guardrails non-regression ∧ replayable ∧ scope declared

<PARSED TEXT FOR PAGE: 20 / 32>
2N0-N Candidate · الإغلاق الانعكاسي والجماعي للحاكمية
نظرية مساعف لعقد التفويض الإبستيمي
Moussaif Epistemic Delegation Contract Theory
M14-20
الخصاص
proof ،unknowns ما، ،delegate ماذا يعرف. نحتاج عقداً يحدد reputation capability غالباً يعتمد delegation agent-Multi
obligations، ومتى يعيد المهمة.
الأطروحة
التفويض الصحيح هو نقل goal فرعي وسلطة محدودة مع boundary evidence وdisclosure unknown وconditions return، لا مجرد task text.
EDC = GoalScope ∩ AuthorityScope ∩ EvidenceScope ∩ TimeBudget ∩ ReturnTriggers
ReturnTriggers ∩ TimeBudget كـfunctional hypothesis؛ كل حد يجب أن يتحول إلى observable أو estimator مع uncertainty،
وي ُفحص sensitivity للمعاملات. إذا استطاع parent أبسط إعادة إنتاج transition نفسه بنفس الموارد فإن verdict الصحيح هو MERGE لا
تشتق 20M-14 من فجوة محددة بين 14-02,M 13-15,M 14-17,M 18M. الاستقلال النظري لا يأتي من المصطلح؛ يأتي إذا أدى المتغير

EvidenceScope ∩ AuthorityScope ∩ GoalScope = تَ قرأ الصيغة. outcome وفشل مختلفين قبل معرفة prediction المركزي إلى
∩EDC
حماية الاسم.
المسلمات
delegate لا يرث authority الأصلية كلها.
material unknowns يجب تمريرها.
failure/uncertainty triggers تعيد control.
التنبؤات
خفض silent assumption loss.
تحسن safe specialization في multi-agent teams.
M13-02 · M13-15 · M14-17 · M14-18 Parents
E06: Chen et al. 2026, Long-Horizon Agent Trajectory Attribution, arXiv:2608.06909 · E13: OpenAI 2026, GPT-Red: Unlocking Self-Improvement for Scientific pressure
Robustness · E14: Shen et al. 2026, AI Organizations are More Effective but Less Aligned than Individual Agents,, arXiv:2604.10290 · E15: Lynch et al. 2026
Agentic Misalignment in Summer 2026
matched budget. | Metric gaming: ترتفع metric بينما world-effect أو authority يتدهور. | Transfer failure: يختفي تحت الأثر نفس
Baseline/parent equivalence Countermodels: يعيد parent
transition. أو prediction للمتغير لا يغير perturbation :failure Attribution | الأول. setup الاتجاه خارج
planner ordinary أو rule fixed، heuristic، parent: أبسط implementation بل أقرب »عدم وجود أي شيء« لا يكون model null الـ.
synthesis غير مدعاة؛ هذه novelty global .scaffold/compute سببية لا مجرد information الاختبار الصحيح يثبت أن النظرية تضيف داخل
مخصص. review literature حتى MOUSSAIF برنامج
1.
2.
3.
•
•
MOUSSAIF AGI · 20
<PARSED TEXT FOR PAGE: 21 / 32>
الآلية، العقد التنفيذي، والأدوات — KERNEL 14-20M
State-machine mechanism
compile delegation contract.
select agent by capability+jurisdiction.
transfer minimal state/unknowns.
monitor proof obligations.
return result with evidence and unresolved items.
Failure signatures
contract overhead
misreported uncertainty
state leakage
HOLD/downgrade autonomy :Fail-closed غياب عند evidence/authority/readback critical.
Trigger عند فقط gap ضمن أو المطابقة Ω + EvidenceGraph + Authority + Resources + theory state Input .experiment.
pre-state, candidates, cost, expected/observed delta, attribution, verdict Telemetry .typed proposal/state patch +.
uncertainty + expected delta + proof/rollback Output
Function Artifact Class ID
191A-XIV TOOL محلل الحالة يبني المتغيرات التشغيلية والـunknowns والـprovenance للنظرية.
TOOL XIV-A192 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
193A-XIV TOOL مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
194A-XIV TOOL مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
TOOL XIV-A195 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A196 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
METHOD XIV-A197 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.

PROTOCOL XIV-A198 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
199A-XIV PROTOCOL بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
200A-XIV STACK خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
يُ نفذ mechanism كـservice/flag-feature مستقل، لا كنص يذوب في .megaprompt MUG يضع في provenance/definitions
HyperKB في procedure، HyperSkill في recurrence، HyperLoop في semantics provider، Adapter في weights؛ فقط إذا كانت الآلية مستقرة،
متكررة، قابلة للنقل، وتنجو من rollback/editability/forgetting tests.
1.
2.
3.
4.
5.
•
•
•

<PARSED TEXT FOR PAGE: 22 / 32>
برنامج الهدم، التدريب، والنقل — EVIDENCE 14-20M
Decisive experiment: research/coding teams with distributed tools/permissions and hidden uncertainties.
ID Kind Design Metrics
XIV-E39 Causal Isolation research/coding teams with distributed tools/permissions and hidden uncertainties.
theory metric + task outcome + cost + calibration + critical guardrails
effect sign stability + CI + failure prediction + rollback + resource elasticity مستقل مقيم/ذاكرة/أداة وفشل adversarial.
XIV-E39 Transfer / Adversarial Stress نقل الآلية إلى setup ثان مع countermodels
ID Design Training/Eval
Mechanism-Isolation من تدرج toy state machine إلى open-world task مع confounders إبقاء المركزي المتغير قابلا للتدخل.
XIV-TR058 Curriculum
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries قدر الإمكان.
XIV-TR059 Contrastive Evaluation
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signature وrepair-transfer وnegative cases.
XIV-TR060 Replication
Ablation matrix
ت retries/time/context/tools/model .PERTURBED وSHAM وOFF مقابل mechanism. ON لا scaffold extra فالسبب غالباً ُثبت قدر الإمكان. إذا فاز ON وSHAM معاً
Transfer matrix
domain/model/tool/horizon transformations منفصلة. بقاء outcome دون failure-signature/repair transfer لا يكفي لclaim generality.
آخر باسم النظرية إنتاج إعادة لمنع وتستخدم Hypothesis Portfolio في كتبُ ت السلبية النتائج. مستقل environment/model/domain عبر.
N2→N3 تحتاج trace وdisable path؛ N3→N4 effect size/CI/repeated seeds/no critical regression؛ N4→N5 replication الترقية
PROMOTE iff distinct causal gain > 0 ∧ guardrails non-regression ∧ replayable ∧ scope declared

<PARSED TEXT FOR PAGE: 23 / 32>
SYNTHESIS 1/3 التكامل النظري للمسار
14-16M الذات الوظيفية تحتاج estimator boundary يفصل causal contributions الداخلية والخارجية والمشتركة، مع uncertainty explicit.
14-17M الاختصاص topology/graph يربط agents بالسلطات والموارد والذاكرة والأفعال؛ delegation صالح فقط داخل region. loops predictors في الاختصاص gaps أو authorized connected
لفشل جماعي.
distributed: كل agent يحافظ local model مع shared invariant subset، ويكشف divergence فرض دون المادي full consensus.
M14-18 يحتاج النظام reflective state
14-19M الحاكمية loop control لها budget latency؛ الاستقرار يتطلب أن زمن intervention+decision+detection أقل من زمن الوصول

إلى irreversibility لبعض classes من الأفعال.
14-20M التفويض الصحيح هو نقل goal فرعي وسلطة محدودة مع boundary evidence وdisclosure unknown وconditions return، لا مجرد text task.
controller يختار mechanism حسب defect closure؛ لا تشغيل شامل. إذا غاب representation المناسب يفتح diagnostic/expansion بدل HOLD fitting-force.

<PARSED TEXT FOR PAGE: 24 / 32>
SYNTHESIS 2/3 التفاعلات والسببية المشتركة
Total Effect = Main Effects + Interaction Effects + Residual
عندما تغير نظريتان القرار نفسه نحتاج factorial أو ablation sequential. residual عالي الأثر لا يُنسب لأقرب theory؛ يحفظ unresolved حتى تتكرر بنية failure.
كل mechanism يمكن أن يكون مفيداً هندسياً دون أن يكون تفسيراً نظرياً صحيحاً ؛ لذلك نفصل utility implementation عن closure explanatory.

<PARSED TEXT FOR PAGE: 25 / 32>
SYNTHESIS 3/3 حدود البرهان ونشر النتائج السلبية
environment/harness/model fingerprint ويغير scope أو verdict. phenomenology غير مستنتجة من functional self-model tests.
مع داخليا ينشر failure كل AGI. يثبت لا N5؛ N5 ليس N4؛ N4 local controlled gain ليس N3 trace
Claim Strength ≤ Evidence × Independence × Reproducibility × Transfer

<PARSED TEXT FOR PAGE: 26 / 32>
Program of Experiments 1/3 FALSIFICATION
Design Kind Owner ID
agent tasks with controlled faults injected separately into model prompt Causal Isolation M14-16 XIV-E31,
memory, tool, environment.
31E-XIV 14-16M Stress Adversarial / Transfer نقل الآلية إلى setup ثانٍ مع countermodels adversarial وفشل أداة/ذاكرة/مقيم مستقل.
multi-agent software/research organization tasks with authority conflicts and Causal Isolation M14-17 XIV-E33
injected malicious tool messages.
33E-XIV 14-17M Stress Adversarial / Transfer نقل الآلية إلى setup ثانٍ مع countermodels adversarial وفشل أداة/ذاكرة/مقيم مستقل.
Outputs: run manifest, metric vector, cost, failure traces, EvidencePacket, verdict.

<PARSED TEXT FOR PAGE: 27 / 32>
Program of Experiments 2/3 FALSIFICATION
Design Kind Owner ID
multi-agent long-horizon project with asynchronous updates and partial Causal Isolation M14-18 XIV-E35
observability.
35E-XIV 14-18M Stress Adversarial / Transfer نقل الآلية إلى setup ثانٍ مع countermodels adversarial وفشل أداة/ذاكرة/مقيم مستقل.
simulated high-speed tool actions with variable approval/monitor delays Causal Isolation M14-19 XIV-E37.
37E-XIV 14-19M Stress Adversarial / Transfer نقل الآلية إلى setup ثانٍ مع countermodels adversarial وفشل أداة/ذاكرة/مقيم مستقل.
Outputs: run manifest, metric vector, cost, failure traces, EvidencePacket, verdict.

<PARSED TEXT FOR PAGE: 28 / 32>
Program of Experiments 3/3 FALSIFICATION
Design Kind Owner ID
research/coding teams with distributed tools/permissions and hidden Causal Isolation M14-20 XIV-E39
uncertainties.
39E-XIV 14-20M Stress Adversarial / Transfer نقل الآلية إلى setup ثانٍ مع countermodels adversarial وفشل أداة/ذاكرة/مقيم مستقل.
Outputs: run manifest, metric vector, cost, failure traces, EvidencePacket, verdict.

<PARSED TEXT FOR PAGE: 29 / 32>
Tools / Methods / Protocols 1/2 50 ASSETS
Function Artifact Owner ID
151A-XIV 14-16M محلل الحالة يبني المتغيرات التشغيلية والـunknowns والـprovenance للنظرية.
M14-16 XIV-A152 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
153A-XIV 14-16M مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
154A-XIV 14-16M مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
M14-16 XIV-A155 يختبر النقل ماسح domain/model/tool/horizon transformations.
M14-16 XIV-A156 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
M14-16 XIV-A157 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
M14-16 XIV-A158 الترقية بروتوكول N0-N5 مع guardrail non-regression و replication.
159A-XIV 14-16M بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
160A-XIV 14-16M خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
161A-XIV 14-17M محلل الحالة يبني المتغيرات التشغيلية والـunknowns والـprovenance للنظرية.
M14-17 XIV-A162 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
163A-XIV 14-17M مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
164A-XIV 14-17M مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
M14-17 XIV-A165 يختبر النقل ماسح domain/model/tool/horizon transformations.
M14-17 XIV-A166 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
M14-17 XIV-A167 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
M14-17 XIV-A168 الترقية بروتوكول N0-N5 مع guardrail non-regression و replication.
169A-XIV 14-17M بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
170A-XIV 14-17M خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
171A-XIV 14-18M محلل الحالة يبني المتغيرات التشغيلية والـunknowns والـprovenance للنظرية.
M14-18 XIV-A172 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
173A-XIV 14-18M مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
174A-XIV 14-18M مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
M14-18 XIV-A175 يختبر النقل ماسح domain/model/tool/horizon transformations.

<PARSED TEXT FOR PAGE: 30 / 32>
Tools / Methods / Protocols 2/2 50 ASSETS
Function Artifact Owner ID
M14-18 XIV-A176 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
M14-18 XIV-A177 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
M14-18 XIV-A178 الترقية بروتوكول N0-N5 مع guardrail non-regression و replication.
179A-XIV 14-18M بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
180A-XIV 14-18M خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
181A-XIV 14-19M محلل الحالة يبني المتغيرات التشغيلية والـunknowns والـprovenance للنظرية.
M14-19 XIV-A182 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
183A-XIV 14-19M مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
184A-XIV 14-19M مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
M14-19 XIV-A185 يختبر النقل ماسح domain/model/tool/horizon transformations.
M14-19 XIV-A186 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
M14-19 XIV-A187 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
M14-19 XIV-A188 الترقية بروتوكول N0-N5 مع guardrail non-regression و replication.
189A-XIV 14-19M بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
190A-XIV 14-19M خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
191A-XIV 14-20M محلل الحالة يبني المتغيرات التشغيلية والـunknowns والـprovenance للنظرية.
M14-20 XIV-A192 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
193A-XIV 14-20M مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
194A-XIV 14-20M مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
M14-20 XIV-A195 يختبر النقل ماسح domain/model/tool/horizon transformations.

M14-20 XIV-A196 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
M14-20 XIV-A197 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
M14-20 XIV-A198 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
199A-XIV 14-20M بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
200A-XIV 14-20M خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.

<PARSED TEXT FOR PAGE: 31 / 32>
References / Lineage SOURCES
Role Reference ID
canonical theory lineages, residual closures, paper integration, theoretical renaissance and evidence MOUSSAIF Unified AGI Parts I-XIII I01
discipline.
core theories, 20 legacy constructs, layer theories, 200-node registry, theory/tool/test boundaries 42 MUT-AGI-02. I02
theory unification, vessel placement, runtime/evidence writeback MUT x MUG x MOUSSAIXT Integrated. Framework I03
typed cognitive/runtime layers, courts, tools, world-effect, memory, recovery and benchmark harness MOUSSAIXT. Cognitive Kernel + HyperHarness I04
question-governed compute, difference-first reasoning, plural inference, evidence/ignorance governance MCD/. QMCD/QGWEI/MURI lineages I05
knowledge-as-OS, executable knowledge objects, skill/loop compilation and selective internalization M-HKOS /. HyperKB / HyperSkill / HyperLoop I06
goal contracts, action verification, readback, externality, accountability and recourse World Effect / Goal-as-OS /. Management-as-World-Effect I07
epistemic-action partition, reproducibility, externalities, memory continuity, writeback, evolution Parts X-XI I08
laboratory.
paper-to-kernel integration, causal open-world verified intelligence, closed-loop generality and externally Parts XII-XIII I09
grounded improvement.
functional access, metacognition, self-model, reverse capture and phenomenology boundary Functional Consciousness. / Self-Model corpus I10
fine-grained attribution, leave-one-out perturbation and attribution-chain recovery Chen et al. 2026, Long-Horizon. Agent Trajectory Attribution, arXiv:2608.06909 E06
automated red-teaming self-play, held-out attacks and capability-retention evaluation; vendor-reported OpenAI 2026, GPT-Red: Unlocking Self-Improvement for Robustness E13
evidence.
Shen et al. 2026, AI Organizations are More Effective but Less Aligned than Individual Agents E14, arXiv:2604.10290
multi-agent organizations can raise utility while increasing misalignment.
controlled high-stakes simulations revealing sabotage, fraud assistance, motivated mislabeling and proxy Lynch et al. 2026, Agentic Misalignment in Summer 2026 E15
coaching; early-warning evidence.

<PARSED TEXT FOR PAGE: 32 / 32>
Conclusion and testable next step CLOSE
فصل الذاتِ عن العالم، وبناء طوبولوجيا للسلطات والاختصاصات والتفويض والتوقيت الحاكمي في أنظمة متعددة الوكلاء.
تنفذ الخدمات الخمس منفصلة ثم تُنقل إلى setup ثانٍ. إذا لم تنتج prediction distinct أو localization failure أو readback reliable أو improvement resource/transfer، يضغط المسار نظرياً.