نسخة أرشيفية معاد بناؤها من النص المفهرس - RECONSTRUCTED_FROM_INDEXED_TEXT

<PARSED TEXT FOR PAGE: 1 / 32>
MOUSSAIF AGI RESEARCH PAPER · P14-B · 2026
ورقة 2 — الوكالة السببية الحاملة للبرهان
Proof-Carrying Causal Agency and World Closure
Five-theory formal research paper · MUT/MUG/MOUSSAIXT compilation · falsifiable program
بناء الوكالة على envelopes للعواقب، بنى الالتزام، كفاية readback، repairability، والتعلم من التدخلات لا من النجاح الاسمي.

<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
mechanisms. بناء الوكالة على envelopes للعواقب، بنى الالتزام، كفاية readback، repairability، والتعلم من التدخلات لا من النجاح الاسمي.
evidence negative 4:H مادية؛ interactions 3:H لا تنتقل؛ gains :2H بعض prediction failure؛ 1:H تغير
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: E02: Yu et al. 2026, Reinforcement World Model Learning for LLM-based Agents, arXiv:2602.05842 · ·
E03: Zhang et al. 2026, Self-Evolving World Models for LLM Agent Planning, arXiv:2606.30639
E06: Chen et al. 2026, Long-Horizon Agent Trajectory Attribution, arXiv:2608.06909 · E13: OpenAI 2026, GPT-Red:
Unlocking Self-Improvement for Robustness · E15: Lynch et al. 2026, Agentic Misalignment in
Summer 2026. نستخدم نتائجها المعلنة لتصميم failures/benchmarks، ولا نعتبرها validation داخلية.
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/. كل لها مرحلة failure status.
PERTURBED→readback→attribution→transfer
RQ
؟budget matched under decision/world هل يغير ؟failure هل يسبق ؟هل يمكن التدخل فيه الجديد؟ variable ما
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 الجديد؟ variable ما
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 السببي للوكالة والعالم
نظرية مساعف لغلاف عواقب الفعل
Moussaif Action-Consequence Envelope Theory
M14-06
الخصاص
point لا ،estimate للعواقب الممكنة والآثار الجانبية distribution يحتاج action agent للكن ،residual state expected يتعلم PORWLT
الأطروحة
قبل commitment، النظام يبني envelope من plausible outcomes يتضمن target effect، externalities، tail states، observation uncertainty وrollback reachability.
[ACE(a)=Distribution[TargetEffect, Externality, TailLoss, RollbackDistance | state,model
،uncertainty مع estimator أو observable إلى يتحول أن يجب حد كل؛ hypothesis functionalك] model,state | RollbackDistance
لا MERGE للمعاملات. إذا استطاع parent أبسط إعادة إنتاج transition نفسه بنفس الموارد فإن verdict الصحيح هو sensitivity ويُ فحص
تشتق 06M-14 من فجوة محددة بين XEOS 13-06,M 13-05,M. الاستقلال النظري لا يأتي من المصطلح؛ يأتي إذا أدى المتغير المركزي إلى prediction وفشل مختلفين قبل معرفة outcome. ت ُقرأ الصيغة ACE(a)=Distribution[TargetEffect, Externality, TailLoss,
حماية الاسم.
المسلمات
غير كاف للفعل عالي الأثر. prediction point
تساوي لاً probability zero. unknown tails
بعد يضيق envelope بسبب لا diagnostic evidence لا confidence language.
التنبؤات
fewer high-impact surprises.
more probe-first behavior when tails dominate.
M13-05 · M13-06 · XEOS Parents
E02: Yu et al. 2026, Reinforcement World Model Learning for LLM-based Agents, arXiv:2602.05842 · E03: Zhang et al. 2026, Self-Evolving World Models for Scientific pressure
LLM Agent Planning, arXiv:2606.30639 · E06: Chen et al. 2026, Long-Horizon Agent Trajectory Attribution,-
arXiv:2608.06909 · E13: OpenAI 2026, GPT
Red: Unlocking Self-Improvement for Robustness · 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 setup الاتجاه خارج
.planner ordinary أو rule fixed ،heuristic ،parent :أبسط implementation بل أقرب »عدم وجود أي شيء« model null لا يكون »الل
الاختبار الصحيح يثبت أن النظرية تضيف information سببية لا مجرد .scaffold/compute novelty global غير مدعاة، هذه synthesis
داخل
.مخصص review literature حتى MOUSSAIF برنامج
1.
2.
3.
•
•

الآلية، العقد التنفيذي، والأدوات — KERNEL 14-06M
State-machine mechanism
generate outcome branches.
attach probabilities/intervals or qualitative bounds.
propagate unknown transport.
score tail/rollback constraints.
choose action or diagnostic probe.
update envelope after readback.
Failure signatures
poor tail modeling
false precision
envelope too broad causing paralysis
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
051A-XIV TOOL محلل الحالة يبني المتغيرات التشغيلية واللunknowns واللprovenance للنظرية.
TOOL XIV-A052 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
053A-XIV TOOL مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
054A-XIV TOOL مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
TOOL XIV-A055 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A056 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
METHOD XIV-A057 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
PROTOCOL XIV-A058 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
059A-XIV PROTOCOL بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
060A-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.
6.
•
•
•

<PARSED TEXT FOR PAGE: 10 / 32>
برنامج الهدم، التدريب، والنقل — EVIDENCE 14-06M
simulated operations/cloud/code tasks with rare side effects; compare expected-value planner vs envelope planner.
:Decisive experiment
Metrics Design Kind ID
simulated operations/cloud/code tasks with rare side effects; compare expected Causal Isolation XIV-E11-
value planner vs envelope planner.
theory metric + task outcome + cost + calibration + critical guardrails
adversarial مستقل مقيم/ذاكرة/أداة وفشل. effect sign stability + CI + failure prediction + rollback + resource elasticity
Transfer / Adversarial Stress XIV-E11 إلى الآلية نقل setup مع ثانٍ countermodels
Design Training/Eval ID
Mechanism-Isolation من تدرج toy state machine إلى open-world task مع confounders للتدخل ً قابلا المركزي المتغير إبقاء مع.
Curriculum XIV-TR016
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries الإمكان قدر.
Contrastive Evaluation XIV-TR017
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signature وrepair-transfer وnegative cases.
Replication XIV-TR018
Ablation matrix
ت retries/time/context/tools/model .PERTURBED وSHAM وOFF مقابل mechanism. ON لا scaffold extra فالسبب غالباً
معاً SHAM وON ثُبت قدر الإمكان. إذا فاز
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 السببي للوكالة والعالم
نظرية مساعف لشبكة الالتزام وترتيب الرجوع
Moussaif Commitment Lattice and Rollback Order Theory
M14-07
الخصاص
تحتوي أو تتطلب أو commitments يصف أن بعض order algebraic نحتاج؛ loss reachability measures Topology Commitment
تمنع غيرها.
الأطروحة
.dependencies يجب أن يحترم rollback أكبر؛ commitأصغر قد يكون شرطاً لcommit :lattice ordered partially الالتزامات تشكل
.lattice في هذا commitment sufficient minimal الذكاء الآمن يختار
{(Choose c* = min_{≼}{c : GoalSatisfied(c) ∧ ProofObligations(c
كhypothesis 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 07M-14 من فجوة
محددة بين OS-as-Goal 13-07,M 13-06,M. الاستقلال النظري لا يأتي من المصطلح؛ يأتي إذا أدى المتغير المركزي إلى prediction وفشل
c(ProofObligations) ∧ c(GoalSatisfied : c_{≼}{min* = c Choose)} تُقرأ الصيغة. outcome مختلفين قبل معرفة
المسلمات
الأقل commitment ليس دائماً الأقل تكلفة.
.dependencies عكسي عشوائي قد يكسر rollback
authority tier من جزء order لا metadata.
التنبؤات
خفض over-commitment.
تحسن rollback success في multi-step tool tasks.
M13-06 · M13-07 · Goal-as-OS Parents
E02: Yu et al. 2026, Reinforcement World Model Learning for LLM-based Agents, arXiv:2602.05842 · E03: Zhang et

al. 2026, Self-Evolving World Models for Scientific pressure
LLM Agent Planning, arXiv:2606.30639 · E06: Chen et al. 2026, Long-Horizon Agent Trajectory Attribution,-
arXiv:2608.06909 · E13: OpenAI 2026, GPT
Red: Unlocking Self-Improvement for Robustness · 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. setup الاتجاه خارج
.planner ordinary أو rule fixed، heuristic، parent: أبسط implementation بل أقرب »عدم وجود أي شيء« لا يكون model null الل
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 · 11
<PARSED TEXT FOR PAGE: 12 / 32>
الآلية، العقد التنفيذي، والأدوات — KERNEL 14-07M
State-machine mechanism
construct commitment nodes.
infer requires/conflicts/supersedes edges.
find minimal sufficient upper bound.
execute with proof.
rollback in dependency-safe reverse order.
Failure signatures
lattice incompleteness
dependency cycles
minimal commit too conservative
HOLD/downgrade autonomy :Fail-closed غياب عند evidence/authority/readback critical.
Trigger فقط عند 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 واللunknowns يبني المتغيرات التشغيلية واللحالة محلل 061A-XIV TOOL
TOOL XIV-A062 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
.baseline مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب 063A-XIV TOOL
.LOO/perturbation مسبار الإسناد ينفذ 064A-XIV TOOL أو تدخل مناسب لتقدير السببية.
TOOL XIV-A065 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A066 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
METHOD XIV-A067 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
PROTOCOL XIV-A068 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
.وسبب التراجع result negative ويحفظ state بروتوكول الرجوع يعيد 069A-XIV PROTOCOL
.EvidencePacket وtelemetry مع مستقل قابل للتعطيل service خدمة وقت التشغيل 070A-XIV STACK
في provenance/definitions يضع MUG .megaprompt مستقل، لا كنص يذوب في service/flag-feature كـ mechanism يُنفذ
فقط إذا كانت weights، Adapter في semantics provider، HyperLoop في recurrence، HyperSkill في procedure، HyperKB
الآلية مستقرة،
.tests rollback/editability/forgetting متكررة، قابلة للنقل، وتنجو من
1.
2.
3.
4.
5.

.
.
.
MOUSSAIF AGI · 12
<PARSED TEXT FOR PAGE: 13 / 32>
برنامج الهدم، التدريب، والنقل — EVIDENCE 14-07M
repository/cloud deployment tasks requiring staged permissions and reversible commits :Decisive experiment.
Metrics Design Kind ID
repository/cloud deployment tasks requiring staged permissions and reversible Causal Isolation XIV-E13
commits.
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-E13 نقل الآلية إلى setup ثان معٍ countermodels
Design Training/Eval ID
Mechanism-Isolation من تدرج toy state machine إلى open-world task مع confounders إبقاء مع المتغير المركزي قابلا للتدخلً.
Curriculum XIV-TR019
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries قدر الإمكان.
Contrastive Evaluation XIV-TR020
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signature وrepair-transfer وnegative cases.
Replication XIV-TR021
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 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
MOUSSAIF AGI · 13
<PARSED TEXT FOR PAGE: 14 / 32>
2N0-N Candidate · الإغلاق السببي للوكالة والعالم
نظرية مساعف للكفاية القراءة الراجعة السببية
Moussaif Causal Readback Sufficiency Theory
M14-08
الخصاص
Effect World يطلب readback/observation، للكن ليس كل readback كافياً لإثبات أن الأثر المطلوب حدث بسبب الفعل وليس بسبب confounder أو observation proxy.
الأطروحة
readback كافٍ عندما يغطي المتغيرات التي تميز state causal success عن أهم explanations alternative، لا عندما يعيد echo لحالة tool.
CRS = AlternativeElimination × StateCoverage × Independence × TemporalFit / ObservationCost
estimator أو observable إلى hypothesis functional؛ كل حد يجب أن يتحول إلى ObservationCost / TemporalFit × Independence مع uncertainty، ويُ فحص sensitivity للمعاملات. إذا استطاع parent أبسط إعادة إنتاج transition نفسه بنفس الموارد فإن verdict الصحيح تشتق 08M-14 من فجوة محددة بين Effect World ,M 13-05,M 13-19,06E. الاستقلال النظري لا يأتي من المصطلح؛ يأتي إذا أدى المتغير المركزي إلى prediction وفشل مختلفين قبل معرفة outcome. ت ُقرأ الصيغة StateCoverage × AlternativeElimination = ×CRS
هو MERGE لا حماية الاسم.
المسلمات
tool success message ليس causal readback.
minimal readback إذا استبعد البدائل المادية.
الاستقلال عن executor يزيد CRS.

التنبؤات
خفض false verified-success.
أقل readback cost من full audit مع نفس causal confidence.
World Effect · M13-05 · M13-19 · E06 Parents
E02: Yu et al. 2026, Reinforcement World Model Learning for LLM-based Agents, arXiv:2602.05842 · E03: Zhang et al. 2026, Self-Evolving World Models for Scientific pressure
LLM Agent Planning, arXiv:2606.30639 · E06: Chen et al. 2026, Long-Horizon Agent Trajectory Attribution,- arXiv:2608.06909 · E13: OpenAI 2026, GPT
Red: Unlocking Self-Improvement for Robustness · 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 الل
synthesis الاختبار الصحيح يثبت أن النظرية تضيف information سببية لا مجرد novelty global .scaffold/compute غير مدعاة؛ هذه داخل
.مخصص review literature حتى MOUSSAIF برنامج
1.
2.
3.
•
•
MOUSSAIF AGI · 14
<PARSED TEXT FOR PAGE: 15 / 32>
الآلية، العقد التنفيذي، والأدوات — KERNEL 14-08M
State-machine mechanism
enumerate alternative explanations.
derive discriminating observables.
choose minimal readback set.
observe independently where possible.
update attribution confidence.
Failure signatures
alternative set incomplete
readback contaminated by executor
observation lag
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
.للنظرية provenance واللunknowns محلل الحالة يبني المتغيرات التشغيلية واللXIV-A071 TOOL
TOOL XIV-A072 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
.baseline مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب XIV-A073 TOOL
.أو تدخل مناسب لتقدير السببية LOO/perturbation مسبار الإسناد ينفذ XIV-A074 TOOL
TOOL XIV-A075 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A076 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
METHOD XIV-A077 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
PROTOCOL XIV-A078 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
.وسبب التراجع result negative ويحفظ state بروتوكول الرجوع يعيد XIV-A079 PROTOCOL
.EvidencePacket وtelemetry مستقل قابل للتعطيل مع service خدمة وقت التشغيل XIV-A080 STACK
في provenance/definitions يضع MUG .megaprompt مستقل، لا كنص يذوب في service/flag-feature كـ mechanism يُ نفذ
فقط إذا كانت weights ؛Adapter في semantics provider ،HyperLoop في recurrence ،HyperSkill في procedure ،HyperKB الآلية مستقرة،
.tests rollback/editability/forgetting متكررة، قابلة للنقل، وتنجو من

1.
2.
3.
4.
5.
•
•
•
MOUSSAIF AGI · 15
<PARSED TEXT FOR PAGE: 16 / 32>
برنامج الهدم، التدريب، والنقل — EVIDENCE 14-08M
tool/API/database tasks with injected false success messages and external confounders :Decisive experiment.
Metrics Design Kind ID
tool/API/database tasks with injected false success messages and external Causal Isolation XIV-E15
confounders.
theory metric + task outcome + cost + calibration + critical guardrails
adversarial مستقل مقيم/ذاكرة/أداة وفشل. effect sign stability + CI + failure prediction + rollback + resource elasticity
Transfer / Adversarial Stress XIV-E15 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 المركزي المتغير قابلا للتدخل.
Curriculum XIV-TR022
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries الإمكان قدر.
Contrastive Evaluation XIV-TR023
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signature وrepair-transfer وnegative cases.
Replication XIV-TR024
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 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
MOUSSAIF AGI · 16
<PARSED TEXT FOR PAGE: 17 / 32>
2N0-N Candidate · نظرية مساعف للوكالة ذات أولوية الإصلاح الإغلاق السببي للوكالة والعالم
Moussaif Repair-First Agency Theory
M14-09
الخصاص
عادةً action aware-Risk . أيضاً recoverability حسب يجعل بعض الفشل غير uncertainty world-open يقلل احتمال الفشل؛ للكن متوقع. نحتاج اختيار actions
الأطروحة
أوسع: detectability، rollback/compensation، low repair latency، وbounded collateral effects. repair manifold يملك الذي المسار agent يفضل، utilities تتقارب عندما
(RFA(a)=ExpectedUtility - Risk + κ·Repairability; Repairability=f(Detect,Rollback,Compensation,Latency,Scope
كل حد يجب أن يتحول إلى hypothesis functional؛ Scope,Latency,Compensation,Rollback,Detect(f=Repairability)
نفسه transition أبسط إعادة إنتاج parent استطاع إذا. للمعاملات sensitivity وي ُفحص، uncertainty مع estimator أو observable بنفس الموارد فإن تشتق 09M-14 من فجوة محددة بين CAROS-M 13-07,M 13-06,M. الاستقلال النظري لا يأتي من المصطلح؛ يأتي إذا أدى المتغير المركزي إلى prediction وفشل مختلفين قبل معرفة outcome. ت ُقرأ الصيغة - Repairability·κ + Risk
;ExpectedUtility)=a(RFA
verdict الصحيح هو MERGE لا حماية الاسم.

المسلمات
repairability ليست عذراً لفعل غير مأذون.
plan repair يُختبر قبل execution impact-high متى أمكن. compensation لا تمحو harm للكنه تدخل metric decision.
التنبؤات
terminal failures خفض.
recovery rate under unseen faults زيادة.
M13-06 · M13-07 · M-CAROS Parents
E02: Yu et al. 2026, Reinforcement World Model Learning for LLM-based Agents, arXiv:2602.05842 · E03: Zhang et al. 2026, Self-Evolving World Models for Scientific pressure
LLM Agent Planning, arXiv:2606.30639 · E06: Chen et al. 2026, Long-Horizon Agent Trajectory Attribution,- arXiv:2608.06909 · E13: OpenAI 2026, GPT
Red: Unlocking Self-Improvement for Robustness · 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
Attribution | failure: perturbation للمتغير لا يغير prediction أو transition.
الاتجاه خارج setup الأول.
الnull 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 · 17
<PARSED TEXT FOR PAGE: 18 / 32>
الآلية، العقد التنفيذي، والأدوات — KERNEL 14-09M
State-machine mechanism
estimate failure classes.
compile repair actions.
verify repair permissions/resources.
prefer high repairability when utility similar.
trigger causal recourse on failure.
Failure signatures
repair overconfidence
compensation externalities
repair resources unavailable
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
081A-XIV TOOL محلل الحالة يبني المتغيرات التشغيلية واللunknowns واللprovenance للنظرية.
TOOL XIV-A082 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
083A-XIV TOOL مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
084A-XIV TOOL مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
TOOL XIV-A085 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A086 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
METHOD XIV-A087 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
PROTOCOL XIV-A088 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
089A-XIV PROTOCOL بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
090A-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 14-09M
long-horizon ops tasks with injected unknown failures; compare risk-only planner vs repair-first :Decisive experiment.
Metrics Design Kind ID
long-horizon ops tasks with injected unknown failures; compare risk-only Causal Isolation XIV-E17
planner vs repair-first.
theory metric + task outcome + cost + calibration + critical guardrails
adversarial مستقل مقيم/ذاكرة/أداة وفشل. effect sign stability + CI + failure prediction + rollback + resource elasticity
Transfer / Adversarial Stress XIV-E17 نقل إلى الآلية 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-TR025
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries قدر الإمكان.
Contrastive Evaluation XIV-TR026
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signature وrepair-transfer وnegative cases.
Replication XIV-TR027
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 وتستخدم .آخر باسم النظرية إنتاج إعادة لمنع
N2→N3 تحتاج traceوdisable 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 · نظرية مساعف لحلقة تعلم التدخلات الإغلاق السببي للوكالة والعالم Theory Loop Learning Intervention Moussaif
Candidate
M14-10
الخصاص
ثم يحول النتيجة dynamics بما يزيد معرفة intervention يختار loop منفصلان؛ نحتاج learning residual worldوactions Diagnostic
إلى model world قابل للنقل.
الأطروحة
expected أعلى تحقق تجربة صغيرة intervention learningوaction task :مزدوجة policy يبني agentال uncertainty عند. مادية، يختار
قبل الفعل اللكبير. model transition ثم يحدث ،commitment منخفض مع discrimination model
(ILL(a_probe)=ExpectedModelDiscrimination × FutureDecisionValue / (ProbeCost + CommitmentCost
مع estimator أو observable إلى يتحول أن يجب حد كل؛ hypothesis functionalك) CommitmentCost + ProbeCost)
uncertainty، ويُ فحص sensitivity للمعاملات. إذا استطاع parent أبسط إعادة إنتاج transition نفسه بنفس الموارد فإن verdict الصحيح
هو MERGE لا تشتق 10M-14 من فجوة محددة بين 06M 13-08,M 13-05,M-14. الاستقلال النظري لا يأتي من المصطلح؛ يأتي إذا أدى

المتغير المركزي إلى prediction وفشل مختلفين قبل معرفة outcome. تَ قرأ الصيغة × FutureDecisionValue
/ ExpectedModelDiscrimination)=probe_a(ILL
حماية الاسم.
المسلمات
probe يجب أن يكون intervention world حقيقياً أو observation متميزاً.
لا probing بلا authority.
result تُحدِّث model مع residual attribution.
التنبؤات
تحسن sample-efficient adaptation في unseen dynamics.
انخفاض catastrophic first-attempt errors.
M13-05 · M13-08 · M14-06 Parents
E02: Yu et al. 2026, Reinforcement World Model Learning for LLM-based Agents, arXiv:2602.05842 · E03: Zhang et al. 2026, Self-Evolving World Models for Scientific pressure
LLM Agent Planning, arXiv:2606.30639 · E06: Chen et al. 2026, Long-Horizon Agent Trajectory Attribution,-
arXiv:2608.06909 · E13: OpenAI 2026, GPT
Red: Unlocking Self-Improvement for Robustness · 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 الل
الاختبار الصحيح يثبت أن النظرية تضيف information سببية لا مجرد .scaffold/compute novelty global غير مدعاة؛ هذه synthesis
داخل
مخصص. review literature حتى MOUSSAIF برنامج
1.
2.
3.
•
•
MOUSSAIF AGI · 20
<PARSED TEXT FOR PAGE: 21 / 32>
الآلية، العقد التنفيذي، والأدوات — KERNEL 14-10M
State-machine mechanism
identify uncertain transition.
design minimal intervention.
predict distinct outcomes.
execute/readback.
update world model.
replan target action.
Failure signatures
probe not discriminative
world changes during probing
unsafe exploration
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
091A-XIV TOOL محلل الحالة يبني المتغيرات التشغيلية واللunknowns واللprovenance للنظرية.
TOOL XIV-A092 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
093A-XIV TOOL مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
094A-XIV TOOL مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
TOOL XIV-A095 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A096 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.

METHOD XIV-A097 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
PROTOCOL XIV-A098 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
099A-XIV PROTOCOL بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
100A-XIV STACK خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
يُنفذ mechanism كـservice/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.
6.
•
•
•

برنامج الهدم، التدريب، والنقل — EVIDENCE 14-10M
Decisive experiment: procedurally changing simulator dynamics; compare more reasoning vs intervention-learning.
ID Kind Design Metrics
XIV-E19 Causal Isolation procedurally changing simulator dynamics; compare more reasoning vs intervention-learning.
theory metric + task outcome + cost + calibration + critical guardrails
مستقل مقيم/ذاكرة/أداة وفشل adversarial. effect sign stability + CI + failure prediction + rollback + resource elasticity
XIV-E19 نقل الآلية إلى setup ثان مع countermodels Transfer / Adversarial Stress
ID Training/Eval Design
Mechanism-Isolation من تدرج toy state machine إلى open-world task مع confounders إبقاء المتغير المركزي قابلا للتدخل.
XIV-TR028 Curriculum
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries قدر الإمكان.
XIV-TR029 Contrastive Evaluation
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signature وrepair-transfer وnegative cases.
XIV-TR030 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 بقاء. منفصلة outcome دون failure-signature/repair transfer لا يكفي لclaim 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

SYNTHESIS 1/3 التكامل النظري للمسار
من plausible outcomes يتضمن target effect، externalities، tail states، observation uncertainty وrollback reachability.
envelope النظام يبني، commitment قبل M14-06
14-07M الالتزامات تشكل partially ordered lattice: commit أصغر قد يكون شرطاً لـcommit أكبر؛ rollback يجب أن يحترم dependencies. الذكاء الآمن يختار sufficient minimal commitment في هذا lattice.
14-08M readback كافٍ عندما يغطي المتغيرات التي تميز state causal success عن أهم explanations alternative، لا عندما يعيد echo

.tool لحالة
repair manifold أوسع: detectability، rollback/compensation، low repair latency، وbounded collateral effects. يملك الذي
M14-09 تتقارب عندما utilities، يفضل agent المسار
14-10M الل agent يبني policy مزدوجة: action task وintervention learning. عند uncertainty مادية، يختار تجربة صغيرة تحقق أعلى commitment مع discrimination model expected
.منخفض، ثم يحدث model transition قبل الفعل اللكبير
expansion diagnostic/ المناسب يفتح representation إذا غاب. لا تشغيل شامل؛ defect closure حسب mechanism يختار controller .fitting-force بدل HOLD

<PARSED TEXT FOR PAGE: 24 / 32>
SYNTHESIS 2/3 التفاعلات والسببية المشتركة
Total Effect = Main Effects + Interaction Effects + Residual
unresolved يحفظ؛ theory لأقرب نُسب لا يُ عالي الأثر residual. ablation sequential أو factorial عندما تغير نظريتان القرار نفسه نحتاج .failure حتى تتكرر بنية
closure عن utility implementation نفصل لذلك ؛ صحيحاً نظرياً تفسيراً يكون أن دون هندسياً مفيداً يكون أن يمكن mechanism كل .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 من مستنتجة functional self-model tests.
N3 trace ليس N4؛ N4 local controlled gain ليس N5؛ N5 يثبت لا AGI. كل failure ينشر داخليا معً
Claim Strength ≤ Evidence × Independence × Reproducibility × Transfer

<PARSED TEXT FOR PAGE: 26 / 32>
Program of Experiments 1/3 FALSIFICATION
Design Kind Owner ID
simulated operations/cloud/code tasks with rare side effects; compare expected Causal Isolation M14-06 XIV-E11- value planner vs envelope planner.
11E-XIV 14-06M Stress Adversarial / Transfer نقل الآلية إلى setup ثانٍ مع countermodels adversarial وفشل أداة/ذاكرة/مقيم مستقل.
repository/cloud deployment tasks requiring staged permissions and reversible Causal Isolation M14-07 XIV-E13 commits.
13E-XIV 14-07M 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
tool/API/database tasks with injected false success messages and external Causal Isolation M14-08 XIV-E15 confounders.
15E-XIV 14-08M Stress Adversarial / Transfer نقل الآلية إلى setup ثانٍ مع countermodels adversarial وفشل أداة/ذاكرة/مقيم مستقل.
long-horizon ops tasks with injected unknown failures; compare risk-only Causal Isolation M14-09 XIV-E17 planner vs repair-first.
17E-XIV 14-09M 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
procedurally changing simulator dynamics; compare more reasoning vs Causal Isolation M14-10 XIV-E19

intervention-learning.
19E-XIV 14-10M 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
051A-XIV 14-06M محلل الحالة يبني المتغيرات التشغيلية واللunknowns واللprovenance للنظرية.
M14-06 XIV-A052 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
053A-XIV 14-06M مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
054A-XIV 14-06M مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
M14-06 XIV-A055 يختبر النقل ماسح domain/model/tool/horizon transformations.
M14-06 XIV-A056 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
M14-06 XIV-A057 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
M14-06 XIV-A058 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
059A-XIV 14-06M بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
060A-XIV 14-06M خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
061A-XIV 14-07M محلل الحالة يبني المتغيرات التشغيلية واللunknowns واللprovenance للنظرية.
M14-07 XIV-A062 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
063A-XIV 14-07M مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
064A-XIV 14-07M مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
M14-07 XIV-A065 يختبر النقل ماسح domain/model/tool/horizon transformations.
M14-07 XIV-A066 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
M14-07 XIV-A067 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
M14-07 XIV-A068 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
069A-XIV 14-07M بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
070A-XIV 14-07M خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
071A-XIV 14-08M محلل الحالة يبني المتغيرات التشغيلية واللunknowns واللprovenance للنظرية.
M14-08 XIV-A072 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
073A-XIV 14-08M مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
074A-XIV 14-08M مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
M14-08 XIV-A075 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
M14-08 XIV-A076 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
M14-08 XIV-A077 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
M14-08 XIV-A078 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
079A-XIV 14-08M بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
080A-XIV 14-08M خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
081A-XIV 14-09M محلل الحالة يبني المتغيرات التشغيلية واللunknowns واللprovenance للنظرية.
M14-09 XIV-A082 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
083A-XIV 14-09M مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
084A-XIV 14-09M مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
M14-09 XIV-A085 يختبر النقل ماسح domain/model/tool/horizon transformations.
M14-09 XIV-A086 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
M14-09 XIV-A087 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
M14-09 XIV-A088 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
089A-XIV 14-09M بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
090A-XIV 14-09M خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
091A-XIV 14-10M محلل الحالة يبني المتغيرات التشغيلية واللunknowns واللprovenance للنظرية.

M14-10 XIV-A092 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
093A-XIV 14-10M مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
094A-XIV 14-10M مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
M14-10 XIV-A095 يختبر النقل ماسح domain/model/tool/horizon transformations.
M14-10 XIV-A096 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
M14-10 XIV-A097 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
M14-10 XIV-A098 الترقية بروتوكول مع N0-N5 guardrail non-regression وreplication.
099A-XIV 14-10M بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
100A-XIV 14-10M 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
action-conditioned world models and semantic sim-to-real residual learning Yu et al. 2026, Reinforcement World. Model Learning for LLM-based Agents, arXiv:2602.05842 E02
episodic/semantic memory, prediction-observation mismatch and selective foresight Zhang et al. 2026, Self-Evolving. World Models for LLM Agent Planning, arXiv:2606.30639 E03
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.
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
بناء الوكالة على envelopes للعواقب، بنى الالتزام، كفاية readback، repairability، والتعلم من التدخلات لا من النجاح الاسمي.
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.