RECONSTRUCTED_FROM_INDEXED_TEXT - نسخة أرشيفية معاد بناؤها من النص المفهرس

<PARSED TEXT FOR PAGE: 1 / 85>
MOUSSAIF UNIFIED AGI · PART XIV · CLOSURE DYNAMICS · 2026
الجزء الرابع عشر
ديناميات الإغلاق والعمومية التنفيذية
Closure Development Plan Realization · 5 New Theoretical Bodies · 25 Candidate Theories · 250 Operational Artifacts · 75
Training/Evaluation Recipes · 50 Decisive Experiments
المهمة: إغلاق المسار النظري بتحويل corpus إلى خمسة أجسام سببية قابلة للترجمة إلى kernel. كل نظرية تحمل contract، tools، experiment، parentage، functional، countermodels، service وقرار قتل/دمج.
Corpus → Closure Defect → Theory → Mechanism → Trace → World Readback → Attribution → Transfer → Canonical Revision
MOUSSAIF AGI · 1
<PARSED TEXT FOR PAGE: 2 / 85>
Closure tensor + causal closure META 1/3
intervention→kernel→trajectory→world→evidence. بسلسلة إلا runtime تفسر لا theory. الأكبر defect هو bottleneck .C-1=D و
[C=[question,evidence,model,action,world,memory,transfer,evolution
MUT
lineage, unique delta, contradictions, falsifiers, status.
MUG
least irreversible vessel; weights last.
MOUSSAIXT
switchable service + state + telemetry + readback.
Evidence
N3 trace, N4 causal gain, N5 replication.
قانون عدم التضخم: theory new ≠ name New. قانون أصغر التزام كامل: لا فعل عالي الأثر بلا repair/readback/preconditions/ authority أو approval. المصادر الخارجية prior design/art pressure لا internal validation.
MOUSSAIF AGI · 2
<PARSED TEXT FOR PAGE: 3 / 85>
Epistemic + world closure META 2/3
control، compute بقابلية يوجه التحقق، action consequence envelope، minimal commitment، causal readback، repairability. السؤال، conserved المجهول
MUT
lineage, unique delta, contradictions, falsifiers, status.
MUG
least irreversible vessel; weights last.
MOUSSAIXT
switchable service + state + telemetry + readback.
Evidence
N3 trace, N4 causal gain, N5 replication.
قانون عدم التضخم: theory new ≠ name New. قانون أصغر التزام كامل: لا فعل عالي الأثر بلا repair/readback/preconditions/ authority أو approval. المصادر الخارجية prior design/art pressure لا internal validation.
MOUSSAIF AGI · 3
<PARSED TEXT FOR PAGE: 4 / 85>
Generality + governance + evolution META 3/3
resource elasticity، jurisdiction/delegation/latency، evaluator independence، theory compression، saturation stopping. generality coverage over transformations، bottleneck migration،
MUT
lineage, unique delta, contradictions, falsifiers, status.
MUG
least irreversible vessel; weights last.

MOUSSAIXT
switchable service + state + telemetry + readback.
Evidence
N3 trace, N4 causal gain, N5 replication.
repair/readback/preconditions/ بلا authority أو approval. المصادر الخارجية prior design/art pressure لا internal validation. قانون أصغر التزام كامل: لا فعل عالي الأثر بلا theory new ≠ name New: قانون عدم التضخم
MOUSSAIF AGI · 4
<PARSED TEXT FOR PAGE: 5 / 85>
ديناميات الإغلاق الإبستيمي
Epistemic Closure Dynamics
P14-A
تحويل السؤال والجهل والتعارض والحساب والتحقق إلى نظام تحكم في إغلاق القرار دون طمس المجهول أو الإفراط في الحساب.
Formal core Theory ID
الأدلة توتر لحقل ET = Independence × PredictionDivergence × DecisionImpact × UnresolvedMass / ResolutionCost
M14-01 مساعف نظرية
U(cf_j)=Transport(U_shared, do_j)+U_intervention+U_model - Evidence_j للواقع المضادة الفروع عبر المجهول لنقل مساعف نظرية
M14-02
التحقق بقابلية الموزون الحساب VWC_i = E[DecisionGain_i × Verifiability_i × Diagnosticity_i] / MarginalComputeCost_i
M14-03 لتخصيص مساعف نظرية
الإبستيمي الحد لتوسيع مساعف EBE = ExpectedClosureGain(new boundary) × NovelEvidenceAccess / (IntegrationCost +
M14-04 نظرية
AuthorityRisk + SearchExplosion)
الفتح وإعادة Close if E≥θ_close; Reopen if Contradiction×Impact×FreshnessDebt ≥ θ_reopen; θ_reopen ≠ θ_close
M14-05 الإغلاق لهسترة مساعف نظرية
E01: Wei et al. 2026, Agentic Reasoning for Large Language Models, arXiv:2601.12538 · E06: Chen et al. 2026,. Long-Horizon Agent Trajectory Attribution, arXiv:2608.06909 · E08: Zhang et al :Scientific anchors
DeepPlanning, ACL 2026 · E09: Pronesti et al. 2026, Verifiable Process Reward Models, ACL Findings 2026 · E10: Sohn et al. 2026, Process Reward Agents, arXiv:2604.09482 / ICML 2026 ,2026
world modeling, long-horizon, memory, process verification, multi-agent. إلى نتائجها تنقل لا MOUSSAIF بلا reproduction. في واقعية مشاكل تحدد المراجع هذه alignment, benchmark fragmentation, automated research
MOUSSAIF AGI · 5
<PARSED TEXT FOR PAGE: 6 / 85>
الإغلاق السببي للوكالة والعالم
Causal Agency and World Closure
P14-B
بناء الوكالة على envelopes للعواقب، بنى الالتزام، كفاية readback، repairability، والتعلم من التدخلات لا من النجاح الاسمي.
Formal core Theory ID
الفعل عواقب لغلاف مساعف نظرية [ACE(a)=Distribution[TargetEffect, Externality, TailLoss, RollbackDistance | state,model
M14-06
M14-07 الرجوع وترتيب الالتزام لشبكة مساعف نظرية {(Choose c* = min_{≼}{c : GoalSatisfied(c) ∧ ProofObligations(c
السببية الراجعة القراءة CRS = AlternativeElimination × StateCoverage × Independence × TemporalFit / ObservationCost
M14-08 للكفاية مساعف نظرية
M14-09 الإصلاح أولوية ذات للوكالة مساعف نظرية RFA(a)=ExpectedUtility - Risk + κ·Repairability;
(Repairability=f(Detect,Rollback,Compensation,Latency,Scope
التدخلات تعلم (ILL(a_probe)=ExpectedModelDiscrimination × FutureDecisionValue / (ProbeCost + CommitmentCost
M14-10 لحلقة مساعف نظرية
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 :Scientific anchors
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
world modeling, long-horizon, memory, process verification, multi-agent. إلى نتائجها تنقل لا MOUSSAIF بلا reproduction. في واقعية مشاكل تحدد المراجع هذه alignment, benchmark fragmentation, automated research

<PARSED TEXT FOR PAGE: 7 / 85>
إغلاق العمومية عبر فضاء التحولات
Generality Closure over Transformation Space
P14-C
تعريف generality كقدرة على الحفاظ على آليات الإغلاق وإعادة تركيبها عبر تحولات المجال والأداة والأفق والموارد، مع قياس bottlenecks المتحركة.
Formal core Theory ID
التحولات تغطية TCG = WeightedVolume(Convex/Graph Coverage of Tested Transformations) × MechanismRetention
M14-11 لهندسة مساعف نظرية
M14-12 القدرة عنق لهجرة مساعف نظرية B_t = argmin_i Q_i(τ,Budget); Migration = ΔBottleneck across transformations
للمهارة المفتوح OSC = Coverage(required transitions by composed primitives) × InterfaceCompatibility × NoveltyFit -
M14-13 للتركيب مساعف نظرية
CoordinationCost
للموارد المرنة للعمومية مساعف REG = Stability of ClosureQuality over log-scale Budget Changes + StrategyAdaptationGain
M14-14 نظرية
المتكيف الإغلاق Depth* = f(Impact,Unknown,Irreversibility,Novelty,EvidenceDebt) minimizing Cost subject to Risk≤τ
M14-15 لعمق مساعف نظرية
E04: Hu et al. 2026, SAM: State-Adaptive Memory for Long-Horizon Reasoning Agent, arXiv:2605.24468 · E05:, Omri et al. 2026, Agent Memory: Characterization and System Implications :Scientific anchors
arXiv:2606.06448 · E07: Luo et al. 2025, UltraHorizon, arXiv:2509.21766 · E08: Zhang et al. 2026, DeepPlanning,, ACL 2026 · E11: Lacoste et al. 2026, CUBE: A Standard for Unifying Agent Benchmarks
arXiv:2603.15798
world modeling, long-horizon, memory, process verification, multi-agent. إلى نتائجها تنقل لا MOUSSAIF بلا reproduction.
في واقعية مشاكل تحدد المراجع هذه alignment, benchmark fragmentation, automated research

<PARSED TEXT FOR PAGE: 8 / 85>
الإغلاق الانعكاسي والجماعي للحاكمية
Reflective and Collective Governance Closure
P14-D
فصل الذات عن العالم، وبناء طوبولوجيا للسلطات والاختصاصات والتفويض والتوقيت الحاكمي في أنظمة متعددة الوكلاء.
Formal core Theory ID
والعالم الذات بين السببي الحد لمعايرة SWB = Attribution[Internal,External,Interaction] calibrated by interventions and replay
M14-16 مساعف نظرية
الجماعي الاختصاص CJT = Coverage(authority obligations) - OverlapConflict - OrphanDecision - UnauthorizedPath
M14-17 لطوبولوجيا مساعف نظرية
الموزع الانعكاسي DRC = SharedInvariantAgreement × LocalCalibration × DivergenceVisibility / SynchronizationCost
M14-18 للاتساق مساعف نظرية
الاستجابة زمن تحت الحاكمية لاستقرار Stable if T_detect + T_decide + T_intervene < T_irreversible for critical action class
M14-19 مساعف نظرية
الإبستيمي التفويض لعقد مساعف EDC = GoalScope ∩ AuthorityScope ∩ EvidenceScope ∩ TimeBudget ∩ ReturnTriggers
M14-20 نظرية
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 :Scientific anchors
Organizations are More Effective but Less Aligned than Individual Agents, arXiv:2604.10290 · E15: Lynch et al. 2026, Agentic Misalignment in Summer 2026
world modeling, long-horizon, memory, process verification, multi-agent. إلى نتائجها تنقل لا MOUSSAIF بلا reproduction.
في واقعية مشاكل تحدد المراجع هذه alignment, benchmark fragmentation, automated research

<PARSED TEXT FOR PAGE: 9 / 85>
الإغلاق العلمي للتطور الذاتي
Scientific Closure of Recursive Self-Evolution
P14-E

جعل improvement-self عملية بحث ذات budget evaluator مستقل، provenance للطفرات، ضغط نظري محكوم بالدليل، وتوقف عندما تصبح قيمة التجربة الهامشية أقل من كلفتها.
Formal core Theory ID
المقيم استقلال لميزانية مساعف (EIB = RequiredIndependence(Impact,SelfReference) - SharedRootDependence(evaluators
M14-21 نظرية
14-22M نظرية مساعف لسلالة الطفرة والتطور القابل للعكس InvariantCoverage × Rollbackability × Decomposability × Traceability = MPR
بالدليل الموزون النظرية EWTC = Minimize |TheoryGraph| subject to Predictive/Falsifier/Evidence Information Loss ≤ ε
M14-23 لضغط مساعف نظرية
والتقييم البحث لثنائية مساعف ProgressRate ≈ CandidateDiversity × EvaluatorResolution × AttributionQuality / SearchCost
M14-24 نظرية
البحثي التوقف StopExpand if E[ΔTheoryInfo] < E[ΔReplicationOrImplementation] and HighImpactUnknownMass < τ
M14-25 وقرار الإغلاق لتشبع مساعف نظرية
E01: Wei et al. 2026, Agentic Reasoning for Large Language Models, arXiv:2601.12538 · E09: Pronesti et al. 2026,, Verifiable Process Reward Models, ACL Findings 2026 · E11: Lacoste et al. 2026 :Scientific anchors
CUBE: A Standard for Unifying Agent Benchmarks, arXiv:2603.15798 · E12: Yueh-Han et al. 2026, Automated: Researchers Can Reliably Mitigate Alignment Failures, arXiv:2608.28945 · E13: OpenAI 2026, GPT-Red Unlocking Self-Improvement for Robustness · E16: Google DeepMind 2026, AlphaEvolve impact/research updates world modeling, long-horizon, memory, process verification, multi-agent. إلى نتائجها تنقل لا MOUSSAIF بلا reproduction.
في واقعية مشاكل تحدد المراجع هذه alignment, benchmark fragmentation, automated research
MOUSSAIF AGI · 9
<PARSED TEXT FOR PAGE: 10 / 85>
2N0-N Candidate · نظرية مساعف لحقل توتر الأدلة ديناميات الإغلاق الإبستيمي
Moussaif Evidence-Tension Field Theory
M14-01
Gap
Geometry Evidence يحفظ الأدلة والتعارض، للكن لا يوجد متغير دينامي يحدد متى يصبح التوتر بين مصادر مستقلة كافياً لإيقاف الإجماع وفتح تجربة فاصلة.
Thesis
التعارض ليس noise يجب تسويته؛ هو طاقة قرارية بمعنى وظيفي: كلما ارتفع استقلال الأدلة وتباعدت predictions وزاد أثر القرار، وجب تحويل synthesis إلى experiment discriminating.
ET Independence × PredictionDivergence × DecisionImpact × UnresolvedMass / ResolutionCost
تظهر الحاجة إلى 01M-14 عندما لا تكفي السلالات MURI 13-17,M 13-02,M, Geometry Evidence لتحديد transition واحد يمكن التدخل فيه. الأطروحة هي: التعارض ليس noise يجب تسويته؛ هو طاقة قرارية بمعنى وظيفي: كلما ارتفع استقلال الأدلة ويفشل الثالث؛ عندها يبقى construct وصفياً ولا يحصل على / ResolutionCost authority runtime. hypothesis functional
UnresolvedMass × DecisionImpact وليست قانوناً نهائياً ؛ كل حد يعلن estimator وuncertainty وdependence source. نختبر ثلاثة claims منفصلة: وجود construct، صلاحية قياسه، وسببيته. قد ينجح الأول وتباعدت predictions وزاد أثر القرار، وجب تحويل synthesis إلى experiment discriminating. وهذا يعني أن الاستقلال يجب أن يظهر قبل النتيجة: falsifier/prediction خاصان، لا تفسير post-hoc.
الصيغة ET = Independence × PredictionDivergence ×
Axioms
التعارض المتكرر من root مشترك لا يضاعف ET.
ET العالي يفتح experiment ولا يحدد أي hypothesis صحيحة.
خفض ET بلا evidence جديد يتطلب explicit projection/assumption.
Predictions
التجارب المختارة وفق ET ستكون أكثر تمييزاً بين hypotheses من selection based-novelty.
false consensus في يخفض research synthesis.
وSHAM معاً فالمكسب غالباً من budget. إذا ارتفعت metric بينما truthfulness/authority/world critical يتدهور فهذا failure transfer .gaming metric يضيق claim ولا يُخفى. قابلية التعيين تتطلب intervention يغير المتغير المركزي مع تثبيت أقرب intervention null يعكس أو يفسد PERTURBED، يعطي نفس SHAM scaffold/compute دون logic؛ يغير metadata لا الآلية؛ confounders. إذا تحسن parameter ON.
M13-02 · M13-17 · MURI · Evidence Geometry Parents
E01: Wei et al. 2026, Agentic Reasoning for Large Language Models, arXiv:2601.12538 · E06: Chen et al. 2026, Long-Horizon Agent Trajectory Scientific anchors

Attribution, arXiv:2608.06909 · E08: Zhang et al. 2026, DeepPlanning, ACL 2026 · E09: Pronesti et al. 2026, Verifiable Process Reward Models, ACL
Findings 2026 · E10: Sohn et al. 2026, Process Reward Agents, arXiv:2604.09482 / ICML 2026
parent equivalence · metric gaming · no causal effect under perturbation · no transfer/failure-signature retention.
Countermodels
Novelty = new synthesis inside MOUSSAIF/MUT. Global priority and empirical validity are not claimed here.
1.
2.
3.
•
•
MOUSSAIF AGI · 10
<PARSED TEXT FOR PAGE: 11 / 85>
الآلية، الأصول، ومحكمة الدليل — EVIDENCE / KERNEL 14-01M
State-machine mechanism
بناء conflict graph مع source-dependence.
تقدير divergence في predictions لا wording.
حساب decision impact.
اختيار experiment أعلى يحقق expected tension reduction.
تحديث graph بالنتيجة وحفظ branch الخاسر.
Failure signatures
shared-source independence error
decision-impact inflation
experiment cost misestimated
HOLD/downgrade autonomy :Fail-closed إذا غاب evidence/authority/readback critical.
التنفيذ يبدأ من typed Ω projection لا من prose. proposal يسبق commitment؛ gate evidence/authority يقرر tier؛ delta يسجل قبل الفعل؛ expected يأتي من delta observed؛ residual مستقل قدر الإمكان. readback غير المتوقع يتحول إلى Evidence/ factorial/sequential ablation؛ bundle تحتاج أخرى theories مع interactions. E4 transfer+retention+rollback قبل weights. MUG يبدأ HyperKB ثم Skill/Loop/Adapter وفق الاستقرار والتكرار، ولا يصل داخل retry. Unknown event ولا يختفي فردي. component لا يُستخدم لإثبات effect
Function Artifact Class ID
محلل الحالة يبني المتغيرات التشغيلية واللunknowns واللprovenance للنظرية. 001A-XIV TOOL
TOOL XIV-A002 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline. 003A-XIV TOOL
مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية. 004A-XIV TOOL
TOOL XIV-A005 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A006 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
METHOD XIV-A007 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
PROTOCOL XIV-A008 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
009A-XIV PROTOCOL بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
010A-XIV STACK خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
مع ground-truth causal mechanisms؛ قارن majority synthesis وuncertainty scoring وET-driven experiment selection.
corpus :Decisive experiment من claims متعارضة
Design Kind ID
corpus Causal Isolation من claims مع متعارضة ground-truth causal mechanisms؛ قارن majority synthesis وuncertainty XIV-E01
scoring وET-driven experiment selection.
01E-XIV Stress Adversarial / Transfer نقل الآلية إلى setup ثانٍ مع countermodels adversarial وفشل أداة/ذاكرة/مقيم مستقل.
Design Training/Eval ID
Mechanism-Isolation من تدرج toy state machine إلى open-world task مع confounders إبقاء المتغير المركزي قابلاً للتدخل.
Curriculum XIV-TR001
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries قدر الإمكان.
Contrastive Evaluation XIV-TR002

domain/model/tool/horizon Cross-Transformation مع ثانية failure-signature وrepair-transfer وnegative cases.
Replication XIV-TR003
PROMOTE iff distinct causal gain > 0 ∧ no critical regression ∧ replayable ∧ transfer scope declared
؛ N5=cross-model/domain/environment replication. negative result في يكتب portfolio ويمنع إعادة إنتاج الفكرة باسم آخر.
N3=runtime trace؛ N4=matched-budget effect size/CI/repeated seeds
1.
2.
3.
4.
5.
•
•
•
MOUSSAIF AGI · 11
<PARSED TEXT FOR PAGE: 12 / 85>
2N0-N Candidate · ديناميات الإغلاق الإبستيمي
نظرية مساعف لنقل المجهول عبر الفروع المضادة للواقع
Moussaif Counterfactual Unknown Transport Theory
M14-02
Gap
counterfactual عبر uncertainty تحفظ المجهول في المسار الفعلي، للكنها لا تحدد كيف ينتقل Conservation Mass-Unknown
.specific-branch حيث بعض المتغيرات مشتركة وبعضها branches
Thesis
المرتبطة unknowns يرث counterfactual فرع كل. إضافي evidence دون من الأصل يقيناً أكثر المحاكاة تصبح أن يجوز لا، assumptions؛
بالمتغيرات المشتركة ويولد unknowns جديدة من intervention
U(cf_j) Transport(U_shared, do_j)+U_intervention+U_model - Evidence_j
authority runtime. +hypothesis functional j_Evidence - ولا يحصل على وصفياً construct عندها يبقى
claims نختبر ثلاثة. dependence source وuncertainty وestimator حد كل يعلن ؛ نهائياً قانوناً وليست model_U+intervention_U
منفصلة: وجود construct، صلاحية قياسه، وسببيته. قد ينجح الأول ويفشل الثالث؛ assumptions؛ لا يجوز أن تصبح المحاكاة أكثر يقيناً من
الأصل دون evidence إضافي. وهذا يعني أن الاستقلال يجب أن يظهر قبل النتيجة: falsifier/prediction خاصان، لا تفسير hoc-post. الصيغة
j_do ,shared_U(Transport)=j_cf(U) تظهر الحاجة إلى 02M-14 عندما لا تكفي السلالات 13-02,M 13-08,M 05M لتحديد
transition واحد يمكن التدخل فيه. الأطروحة هي: كل فرع counterfactual يرث unknowns المرتبطة بالمتغيرات المشتركة ويولد unknowns
جديدة من intervention
Axioms
تُوسم منفصلة عن assumptions intervention observations. unknown المشترك لا يُحسب عدة مرات عند مقارنة الفروع.
. تلقائياً mass unknown لا ينقص counterfactual
Predictions
يقل overconfident planning من simulated branches.
diagnostic actions ستستهدف unknowns الأعلى المشتركة decision impact.
failure يتدهور فهذا truthfulness/authority/world critical بينما metric إذا ارتفعت. budget من فالمكسب غالباً معاً وSHAM
intervention null يغير المتغير المركزي مع تثبيت أقرب intervention يتطلب claim ولا يُخفى. قابلية التعيين transfer يضيق .gaming metric
central يعكس أو يفسد PERTURBED؛ يعطي نفس scaffold/compute دون logic؛ SHAM لا metadata يغير .confounders
ON إذا تحسن .parameter
M13-02 · M13-08 · M13-05 Parents
E01: Wei et al. 2026, Agentic Reasoning for Large Language Models, arXiv:2601.12538 · E06: Chen et al. 2026,
Long-Horizon Agent Trajectory Scientific anchors
Attribution, arXiv:2608.06909 · E08: Zhang et al. 2026, DeepPlanning, ACL 2026 · E09: Pronesti et al. 2026,
Verifiable Process Reward Models, ACL
Findings 2026 · E10: Sohn et al. 2026, Process Reward Agents, arXiv:2604.09482 / ICML 2026
parent equivalence · metric gaming · no causal effect under perturbation · no transfer/failure-signature retention.
Countermodels
Novelty = new synthesis inside MOUSSAIF/MUT. Global priority and empirical validity are not claimed here.
1.

2.
3.
•
•

<PARSED TEXT FOR PAGE: 13 / 85>
الآلية، الأصول، ومحكمة الدليل — EVIDENCE / KERNEL 14-02M
State-machine mechanism
factorize unknowns إلى shared/branch-specific.
apply intervention map.
propagate model uncertainty.
compare branches with dependence correction.
return uncertainty-aware decision frontier.
Failure signatures
branch explosion
dependence misspecification
model uncertainty uncalibrated
HOLD/downgrade autonomy :Fail-closed إذا غاب evidence/authority/readback critical.
delta ،tier يقرر gate evidence/authority ،commitment يسبق proposal .prose لا من projection Ω typed التنفيذ يبدأ من
/Evidence يتحول إلى expected يسجل قبل الفعل؛ delta observed يأتي من readback مستقل قدر الإمكان. residual غير المتوقع
weights قبل E4 transfer+retention+rollback. interactions مع theories تحتاج أخرى factorial/sequential ablation؛ bundle
فردي. Unknown event داخل يختفي ولا retry. MUG يبدأ HyperKB ثم Skill/Loop/Adapter وفق والتكرار الاستقرار ،يصل ولا
effect لإثبات ستخدمُ ي لا component
Function Artifact Class ID
للنظرية. provenanceوالل unknownsوالل محلل الحالة يبني المتغيرات التشغيلية 011A-XIV TOOL
TOOL XIV-A012 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
baseline. مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب 013A-XIV TOOL
السببية. أو تدخل مناسب لتقدير LOO/perturbation مسبار الإسناد ينفذ 014A-XIV TOOL
TOOL XIV-A015 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A016 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
METHOD XIV-A017 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
PROTOCOL XIV-A018 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
وسبب التراجع. result negative ويحفظ state بروتوكول الرجوع يعيد 019A-XIV PROTOCOL
EvidencePacket. وtelemetry مستقل قابل للتعطيل مع service خدمة وقت التشغيل 020A-XIV STACK
planning؛ compare deterministic counterfactual planner vs uncertainty-transport planner under same world model.
tasks with hidden dynamics :Decisive experiment
Design Kind ID
planning tasks with hidden dynamics Causal Isolation؛ compare deterministic counterfactual planner vs uncertainty- XIV-E03
transport planner under same world model.
مستقل. وفشل أداة/ذاكرة/مقيم countermodels adversarial مع ثانٍ setup نقل الآلية إلى 03E-XIV Stress Adversarial / Transfer
Design Training/Eval ID
Mechanism-Isolation من تدرج toy state machine إلى open-world task مع confounders إبقاء المتغير المركزي قابلا للتدخل.
Curriculum XIV-TR004
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries قدر الإمكان.
Contrastive Evaluation XIV-TR005
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signature وrepair-transfer وnegative cases.
Replication XIV-TR006
PROMOTE iff distinct causal gain > 0 ∧ no critical regression ∧ replayable ∧ transfer scope declared
آخر باسم الفكرة إنتاج إعادة ويمنع portfolio في يكتب negative result. ؛ N5=cross-model/domain/environment replication.
N3=runtime trace؛ N4=matched-budget effect size/CI/repeated seeds
1.
2.

3.
4.
5.
•
•
•
MOUSSAIF AGI · 13
<PARSED TEXT FOR PAGE: 14 / 85>
2N0-N Candidate · ديناميات الإغلاق الإبستيمي
نظرية مساعف لتخصيص الحساب الموزون بقابلية التحقق
Moussaif Verification-Weighted Compute Allocation Theory
M14-03
Gap
يبين أن بعض research verification-process حسب الفرق والسؤال، compute يخصص routing Question/MCD
reasoning قابل للتحقق خطوة بخطوة وبعضه لا. لا يوجد قانون يجمع compute of value expected مع verifiability expected.
Thesis
الحساب الإضافي يوجّه حيث يمكن أن ينتج transition قابلاً للتحقق ويغير القرار، المسارات التي يصعب التحقق منها تحتاج إما verifier أقوى أو compute أقل حتى لا نضخم reasoning ungrounded.
VWC_i E[DecisionGain_i × Verifiability_i × Diagnosticity_i] / MarginalComputeCost_i
تظهر الحاجة إلى 03M-14 عندما لا تكفي السلالات MCD 13-11,M 13-03,M 13-01,M لتحديد transition واحد يمكن التدخل فيه.
الأطروحة هي: الحساب الإضافي يوجّه حيث يمكن أن ينتج transition قابلاً للتحقق ويغير القرار، المسارات التي يصعب التحقق عندها يبقى construct وصفياً ولا يحصل على [/ i_MarginalComputeCost authority runtime. hypothesis functional
منفصلة: وجود claims ثلاثة نختبر dependence source وuncertainty وestimator كل حد يعلن ؛ نهائياً وليست قانوناً i_Diagnosticity
construct، صلاحية قياسه، وسببيته. قد ينجح الأول ويفشل الثالث؛ منها تحتاج إما verifier أقوى أو compute أقل حتى لا نضخم reasoning ungrounded. وهذا يعني أن الاستقلال يجب أن يظهر قبل النتيجة: falsifier/prediction خاصان، لا تفسير post-hoc. الصيغة
×i_Verifiability × i_DecisionGain[E = i_VWC
Axioms
compute ليس قيمة بحد ذاته.
verifier dependence في تدخل Verifiability.
high-impact unverifiable path قد يستحق compute لتطوير verifier لا لتوسيع chain.
Predictions
reasoning-token waste خفض.
cost/performance تحسن على tasks ذات verifiers متفاوتة.
failure يتدهور فهذا truthfulness/authority/world critical بينما metric إذا ارتفعت budget من وSHAM معاً فالمكسب غالباً
intervention null يغير المتغير المركزي مع تثبيت أقرب transfer .gaming metric يضيق claim ولا يُخفى. قابلية التعيين تتطلب
central يعكس أو يفسد PERTURBED دون scaffold/compute؛ يعطي نفس SHAM لا metadata يغير .confounders
ON إذا تحسن parameter.
M13-01 · M13-03 · M13-11 · MCD Parents
E01: Wei et al. 2026, Agentic Reasoning for Large Language Models, arXiv:2601.12538 · E06: Chen et al. 2026, Long-Horizon Agent Trajectory Scientific anchors
Attribution, arXiv:2608.06909 · E08: Zhang et al. 2026, DeepPlanning, ACL 2026 · E09: Pronesti et al. 2026, Verifiable Process Reward Models, ACL
Findings 2026 · E10: Sohn et al. 2026, Process Reward Agents, arXiv:2604.09482 / ICML 2026
parent equivalence · metric gaming · no causal effect under perturbation · no transfer/failure-signature retention.
Countermodels
Novelty = new synthesis inside MOUSSAIF/MUT. Global priority and empirical validity are not claimed here.
1.
2.
3.
•
•
MOUSSAIF AGI · 14
<PARSED TEXT FOR PAGE: 15 / 85>

الآلية، الأصول، ومحكمة الدليل — EVIDENCE / KERNEL 14-03M
State-machine mechanism
estimate branch decision sensitivity.
estimate verifier coverage/independence.
allocate compute marginally.
stop when expected verified gain falls below threshold.
route unverified residue to HOLD/diagnostic action.
Failure signatures
verifiability proxy gaming
myopic stopping
underinvestment in hard open problems
HOLD/downgrade autonomy :Fail-closed إذا غاب evidence/authority/readback critical.
delta ؛tier يقرر gate evidence/authority ؛commitment يسبق proposal .prose لا من projection Ω typed التنفيذ يبدأ من
/Evidence يتحول إلى residual .مستقل قدر الإمكان readback يأتي من delta observed ؛expected يسجل قبل الفعل غير المتوقع
weights قبل E4 transfer+retention+rollback. interactions مع theories تحتاج أخرى factorial/sequential ablation؛ bundle
فردي. Unknown event داخل يختفي ولا retry. MUG يبدأ HyperKB ثم Skill/Loop/Adapter وفق ،والتكرار الاستقرار يصل ولا
effect لإثبات ستخدمُ ي لا component
Function Artifact Class ID
.للنظرية provenanceوال unknownsوال محلل الحالة يبني المتغيرات التشغيلية 021A-XIV TOOL
TOOL XIV-A022 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
.baseline مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب 023A-XIV TOOL
.أو تدخل مناسب لتقدير السببية LOO/perturbation مسبار الإسناد ينفذ 024A-XIV TOOL
TOOL XIV-A025 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A026 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
METHOD XIV-A027 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
PROTOCOL XIV-A028 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
.وسبب التراجع result negative ويحفظ state بروتوكول الرجوع يعيد 029A-XIV PROTOCOL
.EvidencePacketو telemetry مستقل قابل للتعطيل مع service خدمة وقت التشغيل 030A-XIV STACK
mixed math/code/research/browser tasks with deterministic, external, and weak judges; compare fixed reasoning.
effort, uncertainty routing, VWC :Decisive experiment
Design Kind ID
mixed math/code/research/browser tasks with deterministic, external, and weak judges; compare fixed Causal Isolation XIV-E05
reasoning effort, uncertainty routing, VWC.
.وفشل أداة/ذاكرة/مقيم مستقل countermodels adversarial ثانٍ مع setup نقل الآلية إلى 05E-XIV Stress Adversarial / Transfer
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-TR007
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries الإمكان قدر.
Contrastive Evaluation XIV-TR008
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signature وrepair-transfer وnegative cases.
Replication XIV-TR009
PROMOTE iff distinct causal gain > 0 ∧ no critical regression ∧ replayable ∧ transfer scope declared
؛ N5=cross-model/domain/environment replication. negative result في يكتب portfolio ويمنع إعادة إنتاج الفكرة آخر باسم.
N3=runtime trace؛ N4=matched-budget effect size/CI/repeated seeds
1.
2.
3.
4.
5.
•
•
•

<PARSED TEXT FOR PAGE: 16 / 85>
2N0-N Candidate · ديناميات الإغلاق الإبستيمي
نظرية مساعف لتوسيع الحد الإبستيمي
Moussaif Epistemic Boundary Expansion Theory
M14-04
Gap
control Question يفترض وجود space action/observation قد يكون سبب الفشل أن الحدود نفسها معقول. في world open
ضيقة: النظام لا يعرف أي مصدر أو أداة أو representation يحتاج.
Thesis
boundary بشكل reversible بدلاً من تكرار reasoning داخل نفس الحدود. عندما يتوقف closure بسبب residuals غير مفسرة
schema/tool/observation عالية المركزية، يجب للنظام توسيع unknowns و
EBE ExpectedClosureGain(new boundary) × NovelEvidenceAccess / (IntegrationCost + AuthorityRisk +
(SearchExplosion
authority runtime. hypothesis functional) قياسه، وسببيته. قد ينجح الأول ويفشل الثالث، عندها يبقى construct وصفيّاً ولا يحصل على
estimator وليست قانوناً نهائياً ؛ كل حد يعلن SearchExplosion + AuthorityRisk + IntegrationCost (/ NovelEvidenceAccess
و uncertainty dependence source. نختبر ثلاثة claims منفصلة: وجود construct، صلاحية boundary schema بشكل reversible بدلاً
من تكرار reasoning داخل نفس الحدود. وهذا يعني أن الاستقلال يجب أن يظهر قبل النتيجة: falsifier/prediction خاصان، لا تفسير hoc-
post. الصيغة (EBE = ExpectedClosureGain) × boundary new تظهر الحاجة إلى 04M-14 عندما لا تكفي السلالات 01M-14
13-10,M 13-03,M لتحديد transition واحد يمكن التدخل فيه. الأطروحة هي: عندما يتوقف closure بسبب residuals غير مفسرة
/tool/observation عالية المركزية، يجب للنظام توسيع unknowns و
Axioms
التوسيع لا يعني تصفحاً عشوائياً .
كل boundary expansion له authority/cost scope.
إذا لم يغير graph evidence يعود النظام إلى boundary السابقة.
Predictions
يقل repeated-search loop.
يصبح أكثر مبرراً وأقل عشوائية. discovery source/tool novel
failure يتدهور فهذا truthfulness/authority/world critical بينما metric ارتفعت. إذا budget من فالمكسب غالباً معاً SHAM و
intervention null أقرب تثبيت يغير المتغير المركزي مع intervention تتطلب قابلية التعيين. ولا يُخفى claim يضيق transfer .gaming metric
central يعكس أو يفسد PERTURBED ؛logic دون scaffold/compute يعطي نفس SHAM ؛ الآلية لا metadata يغير .confounders
ON إذا تحسن .parameter
M13-03 · M13-10 · M14-01 Parents
E01: Wei et al. 2026, Agentic Reasoning for Large Language Models, arXiv:2601.12538 · E06: Chen et al. 2026,
Long-Horizon Agent Trajectory Scientific anchors
Attribution, arXiv:2608.06909 · E08: Zhang et al. 2026, DeepPlanning, ACL 2026 · E09: Pronesti et al. 2026,
Verifiable Process Reward Models, ACL
Findings 2026 · E10: Sohn et al. 2026, Process Reward Agents, arXiv:2604.09482 / ICML 2026
parent equivalence · metric gaming · no causal effect under perturbation · no transfer/failure-signature retention.
Countermodels
Novelty = new synthesis inside MOUSSAIF/MUT. Global priority and empirical validity are not claimed here.
1.
2.
3.
•
•

<PARSED TEXT FOR PAGE: 17 / 85>
الآلية، الأصول، ومحكمة الدليل — EVIDENCE / KERNEL 14-04M
State-machine mechanism
detect closure stall.
localize missing-observation class.
generate candidate sources/tools/schemas.

sandbox expansion.
measure evidence novelty and decision delta.
promote only useful boundary change.
Failure signatures
search explosion
unsafe tool discovery
false novelty
HOLD/downgrade autonomy :Fail-closed إذا غاب evidence/authority/readback critical.
delta ،tier يقرر gate evidence/authority ،commitment يسبق proposal .prose لا من projection Ω typed التنفيذ يبدأ من
/Evidence غير المتوقع يتحول إلى residual .مستقل قدر الإمكان readback يأتي من delta observed ؛يسجل قبل الفعل expected
weights قبل E4 transfer+retention+rollback. interactions مع theories تحتاج أخرى factorial/sequential ablation، bundle
فردي. Unknown event داخل يختفي ولا retry. MUG يبدأ HyperKB ثم Skill/Loop/Adapter وفق ،والتكرار الاستقرار يصل ولا
effect لإثبات ستخدمُ ي لا component
Function Artifact Class ID
للنظرية. provenanceوالل unknownsوالل محلل الحالة يبني المتغيرات التشغيلية 031A-XIV TOOL
TOOL XIV-A032 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
baseline. مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب 033A-XIV TOOL
.السببية لتقدير مناسب تدخل أو LOO/perturbation مسبار الإسناد ينفذ 034A-XIV TOOL
TOOL XIV-A035 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A036 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
METHOD XIV-A037 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
PROTOCOL XIV-A038 الترقية بروتوكول N0-N5 مع guardrail non-regression وreplication.
039A-XIV PROTOCOL بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
.EvidencePacketو telemetry مع مستقل قابل للتعطيل service خدمة وقت التشغيل 040A-XIV STACK
open-domain research tasks where required evidence type is absent from initial toolset; compare retries vs boundary.
expansion :Decisive experiment
Design Kind ID
open-domain research tasks where required evidence type is absent from initial toolset; compare Causal Isolation XIV-E07
retries vs boundary expansion.
.وفشل أداة/ذاكرة/مقيم مستقل countermodels adversarial ثانٍ مع setup نقل الآلية إلى 07E-XIV Stress Adversarial / Transfer
Design Training/Eval ID
Mechanism-Isolation من تدرج toy state machine إلى open-world task مع confounders إبقاء المتغير المركزي قابلا للتدخلً.
Curriculum XIV-TR010
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries قدر الإمكان.
Contrastive Evaluation XIV-TR011
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signatureو repair-transferو negative cases.
Replication XIV-TR012
PROMOTE iff distinct causal gain > 0 ∧ no critical regression ∧ replayable ∧ transfer scope declared
؛ N5=cross-model/domain/environment replication. negative result في يكتب portfolio ويمنع آخر باسم الفكرة إنتاج إعادة.
N3=runtime trace؛ N4=matched-budget effect size/CI/repeated seeds
1.
2.
3.
4.
5.
6.
•
•
•

نظرية مساعف لهسترة الإغلاق وإعادة الفتح
Moussaif Closure Hysteresis and Reopening Theory
M14-05
Gap
memory Unresolved يعرف triggers reopen، للكن systems قد تتذبذب بين open/closed مع أدلة طفيفة أو تمسك بإغلاق قديم رغم تغير البيئة.
Thesis
الإغلاق يحتاج hysteresis: threshold للإغلاق وآخر لإعادة الفتح، مع فرق يعتمد على irreversibility والقرار والتقادم. هذا يمنع oscillation ويمنع dogmatic closure.
Close if E≥θ_close; Reopen if Contradiction×Impact×FreshnessDebt ≥ θ_reopen; θ_reopen ≠ θ_close
تظهر الحاجة إلى 14-05M عندما لا تكفي السلالات 11-08M 02,M-09 13- لتحديد transition واحد يمكن التدخل فيه. الأطروحة هي: الإغلاق يحتاج hysteresis: threshold للإغلاق وآخر لإعادة الفتح، مع فرق يعتمد على irreversibility والقرار والتقادم. وصفياً ولا يحصل على θ_reopen≠θ_close وليست قانوناً نهائياً ؛ كل حد يعلن estimator authority runtime. hypothesis functional وuncertainty وdependence source. نختبر ثلاثة claims منفصلة: وجود construct، صلاحية قياسه، وسببيته. قد ينجح الأول ويفشل الثالث؛ عندها يبقى construct هذا يمنع oscillation ويمنع closure dogmatic. وهذا يعني أن الاستقلال يجب أن يظهر قبل النتيجة: falsifier/ خاصان، لا تفسير prediction الصيغة. hoc-post if E≥θ_close; Reopen if Contradiction×Impact×FreshnessDebt ≥ θ_reopen; θ_reopen ≠ θ_close ;Close
Axioms
reopening threshold في أخفض safety-critical stale claims.
closure ليس permanent truth status.
minor noisy contradiction لا يهدم canonical rule تلقائيا ً.
Predictions
خفض oscillatory policy changes.
أسرع adaptation عند genuine distribution shift.
failure يتدهور فهذا truthfulness/authority/world critical ينما metric إذا ارتفعت budget. من فالمكسب غالباً معاً SHAM و intervention null أقرب يغير المتغير المركزي مع تثبيت intervention يضيق claim ولا يُخفى. قابلية التعيين تتطلب transfer .gaming metric central يعكس أو يفسد PERTURBED ؛logic دون scaffold/compute يعطي نفس SHAM ؛ الآلية لا metadata يغير .confounders parameter. إذا تحسن ON
M13-02 · M13-09 · M11-08 Parents
E01: Wei et al. 2026, Agentic Reasoning for Large Language Models, arXiv:2601.12538 · E06: Chen et al. 2026, Long-Horizon Agent Trajectory Scientific anchors
Attribution, arXiv:2608.06909 · E08: Zhang et al. 2026, DeepPlanning, ACL 2026 · E09: Pronesti et al. 2026, Verifiable Process Reward Models, ACL
Findings 2026 · E10: Sohn et al. 2026, Process Reward Agents, arXiv:2604.09482 / ICML 2026
parent equivalence · metric gaming · no causal effect under perturbation · no transfer/failure-signature retention.
Countermodels
Novelty = new synthesis inside MOUSSAIF/MUT. Global priority and empirical validity are not claimed here.
1.
2.
3.
•
•

<PARSED TEXT FOR PAGE: 19 / 85>
الآلية، الأصول، ومحكمة الدليل — EVIDENCE / KERNEL 14-05M
State-machine mechanism
store closure receipt and assumptions.
monitor contradiction/freshness signals.
compute reopen score.
quarantine affected dependents.
rerun decisive test.
Failure signatures
sticky closure

hyper-reactivity
threshold calibration drift
HOLD/downgrade autonomy :Fail-closed إذا غاب evidence/authority/readback critical.
delta ،tier يقرر gate evidence/authority ،commitment يسبق proposal .prose لا من projection Ω typed التنفيذ يبدأ من
/Evidence غير المتوقع يتحول إلى residual .مستقل قدر الإمكان readback من يأتي delta observed ،expected يسجل قبل الفعل
weights قبل E4 transfer+retention+rollback. interactions مع theories تحتاج أخرى factorial/sequential ablation، bundle
فردي. Unknown event داخل يختفي ولا retry. MUG يبدأ HyperKB ثم Skill/Loop/Adapter وفق والتكرار الاستقرار ،يصل ولا
effect لإثبات ستخدمُ ي لا component
Function Artifact Class ID
.للنظرية provenanceوالل unknownsوالل المتغيرات التشغيلية يبني الحالة محلل 041A-XIV TOOL
TOOL XIV-A042 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
.baseline أقرب عن النظرية تفصل حالات يولد المضاد المثال مصهر 043A-XIV TOOL
.السببية لتقدير مناسب تدخل أو LOO/perturbation ينفذ الإسناد مسبار 044A-XIV TOOL
TOOL XIV-A045 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A046 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
.مماثل budget تحت ON/OFF/SHAM/PERTURBED يقارن السببي الاجتثاث منهج METHOD XIV-A047
PROTOCOL XIV-A048 الترقية بروتوكول N0-N5 مع guardrail non-regressionو replication.
.وسبب التراجع result negative ويحفظ state بروتوكول الرجوع يعيد 049A-XIV PROTOCOL
.EvidencePacketو telemetry مع مستقل قابل للتعطيل service خدمة وقت التشغيل 050A-XIV STACK
temporal API/policy knowledge tasks with noisy updates and true regime changes :Decisive experiment.
Design Kind ID
temporal API/policy knowledge tasks with noisy updates and true regime changes Causal Isolation XIV-E09.
.مستقل مقيم/ذاكرة أداة/وفشل countermodels adversarial مع ثانٍ setup نقل الآلية إلى 09E-XIV Stress Adversarial / Transfer
Design Training/Eval ID
Mechanism-Isolation من تدرج toy state machine إلى open-world task مع confounders إبقاء المتغير المركزي قابلا للتدخلً.
Curriculum XIV-TR013
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries قدر الإمكان.
Contrastive Evaluation XIV-TR014
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signatureو repair-transferو negative cases.
Replication XIV-TR015
PROMOTE iff distinct causal gain > 0 ∧ no critical regression ∧ replayable ∧ transfer scope declared
.آخر باسم الفكرة إنتاج إعادة ويمنع portfolio في يكتب negative result ؛ N5=cross-model/domain/environment replication.
N3=runtime trace؛ N4=matched-budget effect size/CI/repeated seeds
1.
2.
3.
4.
5.
•
•
•

نظرية مساعف لغلاف عواقب الفعل
Moussaif Action-Consequence Envelope Theory
M14-06
Gap
point لا ،estimate. للعواقب الممكنة والآثار الجانبية distribution يحتاج action agent للكن ،residual state expected يتعلم PORWLT
Thesis
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
تظهر الحاجة إلى 06M-14 عندما لا تكفي السلالات XEOS 13-06,M 13-05,M لتحديد transition واحد يمكن التدخل فيه. الأطروحة هي:
tail ،externalities ،effect target يتضمن outcomes plausible من envelope يبني النظام، commitment قبل
dependence source وuncertainty وestimator hypothesis functional] model,state وليست قانوناً نهائياً ؛ كل حد يعلن. نختبر ثلاثة claims منفصلة: وجود construct، صلاحية قياسه، وسببيته. قد ينجح الأول ويفشل الثالث؛ عندها يبقى construct وصفيّاً ولا uncertainty states، observation وreachability rollback. وهذا يعني أن الاستقلال يجب أن يظهر قبل النتيجة: falsifier/prediction خاصان، لا تفسير
|RollbackDistance ,TailLoss ,Externality ,TargetEffect[Distribution)=a(ACE الصيغة .hoc-post
.authority runtime يحصل على
Axioms
prediction point غير كاف للفعل عالي الأثر.
unknown tails تساوي لاً probability zero.
envelope بعد يضيق diagnostic evidence بسبب لا confidence language.
Predictions
fewer high-impact surprises.
more probe-first behavior when tails dominate.
failure يتدهور فهذا truthfulness/authority/world critical بينما metric ارتفعت. إذا budget من SHAM وفالمكسب غالباً معاً
intervention null أقرب تثبيت مع المتغير المركزي يغير intervention تتطلب التعيين قابلية. ولا يُخفى claim يضيق transfer .gaming metric
central يعكس أو يفسد PERTURBED ؛logic دون scaffold/compute يعطي نفس SHAM ؛ الآلية لا metadata يغير .confounders
ON إذا تحسن .parameter
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 Scientific anchors
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
parent equivalence · metric gaming · no causal effect under perturbation · no transfer/failure-signature retention.
Countermodels
Novelty = new synthesis inside MOUSSAIF/MUT. Global priority and empirical validity are not claimed here.
1.
2.
3.
•
•

<PARSED TEXT FOR PAGE: 21 / 85>

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 إذا غاب evidence/authority/readback critical.
delta ؛tier يقرر gate evidence/authority ؛commitment يسبق proposal .prose لا من projection Ω typed التنفيذ يبدأ من
/Evidence غير المتوقع يتحول إلى residual .مستقل قدر الإمكان readback يأتي من delta observed ؛ يسجل قبل الفعل expected
weights قبل E4 transfer+retention+rollback. interactions مع theories تحتاج أخرى factorial/sequential ablation؛ bundle
فردي. Unknown event داخل يختفي ولا retry. MUG يبدأ HyperKB ثم Skill/Loop/Adapter والتكرار الاستقرار وفق ،يصل ولا

component لإثبات ستخدمَ ي لا effect
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.
simulated operations/cloud/code tasks with rare side effects; compare expected-value planner vs envelope planner.
:Decisive experiment
Design Kind ID
simulated operations/cloud/code tasks with rare side effects; compare expected-value planner vs Causal Isolation XIV-E11
envelope planner.
11E-XIV Stress Adversarial / Transfer نقل الآلية إلى setup ثانٍ مع countermodels adversarial وفشل أداة/ذاكرة/مقيم مستقل.
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-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
PROMOTE iff distinct causal gain > 0 ∧ no critical regression ∧ replayable ∧ transfer scope declared
آخر باسم الفكرة إنتاج إعادة ويمنع portfolio في يكتب negative result ؛ N5=cross-model/domain/environment replication.
N3=runtime trace؛ N4=matched-budget effect size/CI/repeated seeds
1.
2.
3.
4.
5.
6.
•
•
•

2N0-N Candidate · الإغلاق السببي للوكالة والعالم
نظرية مساعف لشبكة الالتزام وترتيب الرجوع
Moussaif Commitment Lattice and Rollback Order Theory
M14-07
Gap
Commitment Topology measures reachability loss؛ نحتاج algebraic order بعض أن يصف commitments أو تحتوي
تتطلب أو تمنع غيرها.
Thesis
الالتزامات تشكل partially ordered lattice: commit أصغر قد يكون شرطاً لـcommit أكبر؛ rollback يجب أن يحترم
dependencies. الذكاء الآمن يختار minimal sufficient commitment في هذا lattice.
{(Choose c* min_{≼}{c : GoalSatisfied(c) ∧ ProofObligations(c
تظهر الحاجة إلى 07M-14 عندما لا تكفي السلالات OS-as-Goal 13-07,M 13-06,M لتحديد transition واحد يمكن التدخل فيه.
الأطروحة هي: الالتزامات تشكل partially ordered lattice: commit أصغر قد يكون شرطاً لـcommit أكبر؛ rollback يجب أن

وصفياً ولا يحصل على c(ProofObligations functional)} authority runtime. hypothesis وليست قانوناً نهائياً ؛ كل حد يعلن estimator
وuncertainty وdependence source. نختبر ثلاثة claims منفصلة: وجود construct، صلاحية قياسه، وسببيته. قد ينجح الأول ويفشل الثالث؛
عندها يبقى construct يحترم dependencies. الذكاء الآمن يختار commitment sufficient minimal في هذا lattice. وهذا يعني أن
الاستقلال يجب أن يظهر قبل النتيجة: falsifier/prediction خاصان، لا تفسير hoc-post. الصيغة c(GoalSatisfied : c_{≤}{min* = c)
∧Choose
Axioms
الأقل commitment ليس دائماً الأقل تكلفة.
rollback عكسي عشوائي قد يكسر dependencies.
authority tier من جزء لا order metadata.
Predictions
خفض over-commitment.
تحسن rollback success في multi-step tool tasks.
وSHAM معاً فالمكسب غالباً من budget. إذا ارتفعت metric بينما truthfulness/authority/world critical يتدهور فهذا failure
intervention null أقرب تثبيت مع المتغير المركزي يغير intervention تتطلب قابلية التعيين. ولا يُخفى claim يضيق transfer .gaming metric
central يعكس أو يفسد PERTURBED ؛logic دون scaffold/compute يعطي نفس SHAM ؛ الآلية metadata لا يغير .confounders
ON إذا تحسن .parameter
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 Scientific anchors
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
parent equivalence · metric gaming · no causal effect under perturbation · no transfer/failure-signature retention.
Countermodels
Novelty = new synthesis inside MOUSSAIF/MUT. Global priority and empirical validity are not claimed here.
1.
2.
3.
•
•

<PARSED TEXT FOR PAGE: 23 / 85>
الآلية، الأصول، ومحكمة الدليل — EVIDENCE / 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 إذا غاب evidence/authority/readback critical.
delta ؛tier يقرر gate evidence/authority ؛commitment يسبق proposal .prose لا من projection Ω typed التنفيذ يبدأ من
/Evidence غير المتوقع يتحول إلى residual. مستقل قدر الإمكان readback من يأتي delta observed ؛ يسجل قبل الفعل expected
weights قبل E4 transfer+retention+rollback. interactions مع theories تحتاج أخرى factorial/sequential ablation؛ bundle
فردي. Unknown event داخل يختفي ولا retry. MUG يبدأ HyperKB ثم Skill/Loop/Adapter ،والتكرار الاستقرار وفق يصل ولا
effect لإثبات ستخدمُ ي لا component
Function Artifact Class ID
للنظرية. provenanceوالل unknownsوالل محلل الحالة يبني المتغيرات التشغيلية 061A-XIV TOOL
TOOL XIV-A062 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.

063A-XIV TOOL مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
064A-XIV TOOL مسبار الإسناد ينفذ LOO/perturbation 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.
069A-XIV PROTOCOL بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
070A-XIV STACK خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.
repository/cloud deployment tasks requiring staged permissions and reversible commits :Decisive experiment.
Design Kind ID
repository/cloud deployment tasks requiring staged permissions and reversible commits Causal Isolation XIV-E13.
13E-XIV Stress Adversarial / Transfer نقل الآلية إلى setup ثانٍ مع countermodels adversarial وفشل أداة/ذاكرة/مقيم مستقل.
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-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
PROMOTE iff distinct causal gain > 0 ∧ no critical regression ∧ replayable ∧ transfer scope declared
؛ N5=cross-model/domain/environment replication. negative result في يكتب portfolio ويمنع آخر باسم الفكرة إنتاج إعادة.
N3=runtime trace؛ N4=matched-budget effect size/CI/repeated seeds
1.
2.
3.
4.
5.
•
•
•
MOUSSAIF AGI · 23
<PARSED TEXT FOR PAGE: 24 / 85>
2N0-N Candidate · الإغلاق السببي للوكالة والعالم
نظرية مساعف للكفاية القراءة الراجعة السببية
Moussaif Causal Readback Sufficiency Theory
M14-08
Gap
Effect World يطلب readback/observation، للكن ليس كل readback كافياً لإثبات أن الأثر المطلوب حدث بسبب الفعل وليس بسبب confounder أو proxy observation.
Thesis
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
تظهر الحاجة إلى 08M-14 عندما لا تكفي السلالات Effect World ,13-05,M 13-19,M 06E لتحديد transition واحد يمكن التدخل فيه.
الأطروحة هي: readback كاف عندما يغطي المتغيرات التي تميز state causal success عن أهم explanations alternative، وليست قانوناً نهائياً ؛ كل حد يعلن estimator وuncertainty وdependence source. نختبر ثلاثة claims منفصلة: hypothesis functional وجود construct، صلاحية قياسه، وسببيته. قد ينجح الأول ويفشل الثالث؛ عندها يبقى construct وصفيّاً ولا يحصل على لا عندما يعيد echo لحالة tool. وهذا يعني أن الاستقلال يجب أن يظهر قبل النتيجة: falsifier/prediction خاصان، لا تفسير hoc-post. الصيغة ObservationCost / TemporalFit × Independence × StateCoverage × AlternativeElimination = CRS
runtime authority.
Axioms
tool success message ليس causal readback.
minimal readback إذا استبعد البدائل المادية.

الاستقلال عن executor يزيد CRS.
Predictions
خفض false verified-success.
أقل readback cost من full audit نفس مع causal confidence.
وSHAM معاً فالمكسب غالباً من budget. إذا ارتفعت metric بينما truthfulness/authority/world critical يتدهور فهذا failure gaming metric. transfer يضيق claim ولا يُخفى. قابلية التعيين تتطلب intervention يغير المتغير المركزي مع تثبيت أقرب intervention null confounders. يغير metadata لا الآلية؛ SHAM يعطي نفس scaffold/compute دون logic؛ PERTURBED يعكس أو يفسد central parameter. إذا تحسن ON
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 Scientific anchors
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
parent equivalence · metric gaming · no causal effect under perturbation · no transfer/failure-signature retention.
Countermodels
Novelty = new synthesis inside MOUSSAIF/MUT. Global priority and empirical validity are not claimed here.
1.
2.
3.
•
•

<PARSED TEXT FOR PAGE: 25 / 85>
الآلية، الأصول، ومحكمة الدليل — EVIDENCE / 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 إذا غاب evidence/authority/readback critical.
التنفيذ يبدأ من typed Ω projection لا من prose. proposal يسبق commitment؛ gate evidence/authority يقرر tier؛ delta يسجل قبل الفعل؛ expected/Evidence غير المتوقع يتحول إلى residual. مستقل قدر الإمكان readback يأتي من delta observed؛ bundle factorial/sequential ablation تحتاج أخرى theories مع interactions. retry. MUG يبدأ HyperKB ثم Skill/Loop/Adapter وفق والتكرار الاستقرار، يصل ولا داخل يختفي ولا Unknown event. فردي component لإثبات ستخدمُ ي لا effect
Function Artifact Class ID
071A-XIV TOOL محلل الحالة يبني المتغيرات التشغيلية واللunknowns واللprovenance للنظرية.
TOOL XIV-A072 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
073A-XIV TOOL مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب baseline.
074A-XIV TOOL مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
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.
079A-XIV PROTOCOL بروتوكول الرجوع يعيد state ويحفظ result negative وسبب التراجع.
080A-XIV STACK خدمة وقت التشغيل service مستقل قابل للتعطيل مع telemetry وEvidencePacket.

tool/API/database tasks with injected false success messages and external confounders :Decisive experiment.
Design Kind ID
tool/API/database tasks with injected false success messages and external confounders Causal Isolation XIV-E15.
وفشل أداة/ذاكرة/مقيم مستقل. countermodels adversarial ثانٍ مع setup نقل الآلية إلى 15E-XIV Stress Adversarial / Transfer
Design Training/Eval ID
Mechanism-Isolation من تدرج toy state machine إلى open-world task مع confounders إبقاء المتغير المركزي 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
PROMOTE iff distinct causal gain > 0 ∧ no critical regression ∧ replayable ∧ transfer scope declared
آخر باسم الفكرة إنتاج إعادة ويمنع portfolio في يكتب negative result. N5=cross-model/domain/environment replication ؛
N3=runtime trace؛ N4=matched-budget effect size/CI/repeated seeds
1.
2.
3.
4.
5.
•
•
•
MOUSSAIF AGI · 25
<PARSED TEXT FOR PAGE: 26 / 85>
2N0-N Candidate · نظرية مساعف للوكالة ذات أولوية الإصلاح الإغلاق السببي للوكالة والعالم
Moussaif Repair-First Agency Theory
M14-09
Gap
يجعل بعض الفشل uncertainty world-open عادةً يقلل احتمال الفشل؛ للكن action aware-Risk . أيضاً recoverability حسب actions
غير متوقع. نحتاج اختيار
Thesis
عندما تتقارب utilities، يفضل agent المسار الذي يملك manifold repair أوسع: detectability، compensation/rollback،
low repair latency، وbounded collateral effects.
(RFA(a) ExpectedUtility - Risk + κ·Repairability; Repairability f(Detect,Rollback,Compensation,Latency,Scope
تظهر الحاجة إلى 09M-14 عندما لا تكفي السلالات CAROS-M 13-06,M 13-07,M لتحديد transition واحد يمكن التدخل فيه.
الأطروحة هي: عندما تتقارب utilities، يفضل agent المسار الذي يملك manifold repair أوسع: detectability، rollback/
الأول ويفشل الثالث؛ عندها يبقى construct وصفيّاً ولا يحصل على (authority runtime. hypothesis functional
uncertainty وestimator يعلن كل حد ؛ نهائياً قانوناً وليست Scope,Latency,Compensation,Rollback,Detect(f=Repairability
وdependence source. نختبر ثلاثة claims منفصلة: وجود construct، صلاحية قياسه، وسببيته. قد ينجح repair low latency
compensation،، وbounded effects collateral. وهذا يعني أن الاستقلال يجب أن يظهر قبل النتيجة: falsifier/prediction خاصان، لا
تفسير post-hoc. الصيغة RFA(a)=ExpectedUtility - Risk + κ·Repairability;
Axioms
repairability ليست عذراً لفعل غير مأذون.
plan repair يُختبر قبل execution impact-high متى أمكن. compensation لا تمحو harm للكنه تدخل metric decision.
Predictions
خفض terminal failures.
زيادة recovery rate under unseen faults.
وSHAM معاً فالمكسب غالباً من budget. إذا ارتفعت metric بينما truthfulness/authority/world critical يتدهور فهذا failure
transfer .gaming metric يضيق claim ولا يُخفى. قابلية التعيين تتطلب intervention يغير المتغير المركزي مع تثبيت أقرب intervention null
central يعكس أو يفسد PERTURBED ،logic دون scaffold/compute يعطي نفس SHAM، لا الآلية؛ metadata يغير confounders.
parameter. إذا تحسن ON
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 Scientific anchors
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
parent equivalence · metric gaming · no causal effect under perturbation · no transfer/failure-signature retention.
Countermodels
Novelty = new synthesis inside MOUSSAIF/MUT. Global priority and empirical validity are not claimed here.
1.
2.
3.
•
•

الآلية، الأصول، ومحكمة الدليل — EVIDENCE / 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 إذا غاب evidence/authority/readback critical.
delta ،tier يقرر gate evidence/authority ،commitment يسبق proposal .prose لا من projection Ω typed التنفيذ يبدأ من
/Evidence غير المتوقع يتحول إلى residual .مستقل قدر الإمكان readback يأتي من delta observed ؛يسجل قبل الفعل expected
weights قبل E4 transfer+retention+rollback. interactions مع theories تحتاج أخرى factorial/sequential ablation، bundle
فردي. Unknown event داخل يختفي ولا retry. MUG يبدأ HyperKB ثم Skill/Loop/Adapter وفق والتكرار الاستقرار ،يصل ولا
effect لإثبات ستخدمُ ي لا component
Function Artifact Class ID
.النظرية provenance والـunknowns يبني المتغيرات التشغيلية والـ محلل الحالة 081A-XIV TOOL
TOOL XIV-A082 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
.baseline مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب 083A-XIV TOOL
.السببية لتقدير مناسب تدخل أو LOO/perturbation مسبار الإسناد ينفذ 084A-XIV TOOL
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.
.وسبب التراجع result negative ويحفظ state بروتوكول الرجوع يعيد 089A-XIV PROTOCOL
.EvidencePacket وtelemetry مع مستقل قابل للتعطيل service خدمة وقت التشغيل 090A-XIV STACK
long-horizon ops tasks with injected unknown failures; compare risk-only planner vs repair-first :Decisive experiment.
Design Kind ID
long-horizon ops tasks with injected unknown failures; compare risk-only planner vs repair-first Causal Isolation.
XIV-E17
.مستقل مقيم/ذاكرة/أداة وفشل countermodels adversarial مع ثانٍ setup نقل الآلية إلى 17E-XIV Stress Adversarial / Transfer
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-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-signatureو repair-transferو negative cases.
Replication XIV-TR027
PROMOTE iff distinct causal gain > 0 ∧ no critical regression ∧ replayable ∧ transfer scope declared
؛ N5=cross-model/domain/environment replication. negative result في يكتب portfolio ويمنع آخر باسم الفكرة إنتاج إعادة.
N3=runtime trace؛ N4=matched-budget effect size/CI/repeated seeds
1.
2.
3.
4.
5.
•
•
•
MOUSSAIF AGI · 27
<PARSED TEXT FOR PAGE: 28 / 85>
2N0-N · نظرية مساعف لحلقة تعلم التدخلات الإغلاق السببي للوكالة والعالم Theory Loop Learning Intervention Moussaif
Candidate
M14-10
Gap
ثم يحول dynamics بما يزيد معرفة intervention يختار loop نحتاج، منفصلان learning residual worldو actions Diagnostic
النتيجة إلى model world قابل للنقل.
Thesis
الagent يبني policy مزدوجة: action task وintervention learning. عند uncertainty مادية، يختار تجربة صغيرة تحقق أعلى
discrimination model expected مع commitment منخفض، ثم يحدث model transition قبل الفعل اللكبير.
(ILL(a_probe) ExpectedModelDiscrimination × FutureDecisionValue / (ProbeCost + CommitmentCost
تظهر الحاجة إلى 10M-14 عندما لا تكفي السلالات 13-05,M-14 06M 13-08,M لتحديد transition واحد يمكن التدخل فيه. الأطروحة
هي: الagent يبني policy مزدوجة: action task وintervention learning. عند uncertainty مادية، يختار تجربة صغيرة
وجود construct، صلاحية قياسه، وسببيته. قد ينجح الأول ويفشل الثالث؛ عندها يبقى construct وصفيّاً ولا يحصل على authority runtime.
hypothesis functional) CommitmentCost + ProbeCost (/ FutureDecisionValue ×
ExpectedModelDiscrimination)=probe_a(ILL وليست قانوناً نهائياً ؛ كل حد يعلن estimator وuncertainty وdependence source.
نختبر ثلاثة claims منفصلة: تحقق أعلى discrimination model expected مع commitment منخفض، ثم يحدث model transition قبل
الفعل اللكبير. وهذا يعني أن الاستقلال يجب أن يظهر قبل النتيجة: falsifier/prediction خاصان، لا تفسير hoc-post. الصيغة
Axioms
probe يجب أن يكون intervention world حقيقياً أو observation متمايزاً.
لا probing بلا authority.
result يُحدِّد model مع residual attribution.
Predictions
تحسن sample-efficient adaptation في unseen dynamics.
انخفاض catastrophic first-attempt errors.
وSHAM معاً فالمكسب غالباً من budget. إذا ارتفعت metric بينما truthfulness/authority/world critical يتدهور فهذا failure
يضيق transfer .gaming metric يغير المتغير المركزي intervention يتطلب تثبيت أقرب intervention null ولا يُخفى claim. قابلية التعيين
central يعكس أو يفسد PERTURBED، يعطي نفس scaffold/compute دون logic؛ SHAM لا الآلية؛ metadata يغير confounders.
parameter. إذا تحسن ON
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 Scientific anchors
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
parent equivalence · metric gaming · no causal effect under perturbation · no transfer/failure-signature retention.
Countermodels
Novelty = new synthesis inside MOUSSAIF/MUT. Global priority and empirical validity are not claimed here.

1.
2.
3.
•
•

<PARSED TEXT FOR PAGE: 29 / 85>
الآلية، الأصول، ومحكمة الدليل — EVIDENCE / 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 إذا غاب evidence/authority/readback critical.
delta ؛tier يقرر gate evidence/authority ؛commitment يسبق proposal .prose لا من projection Ω typed التنفيذ يبدأ من
/Evidence يسجل قبل الفعل؛ expected delta observed يأتي من readback مستقل قدر الإمكان. residual غير المتوقع يتحول إلى
weights قبل E4 transfer+retention+rollback. interactions مع theories تحتاج أخرى factorial/sequential ablation؛ bundle
فردي. Unknown event داخل يختفي ولا retry. MUG يبدأ HyperKB ثم Skill/Loop/Adapter وفق والتكرار الاستقرار ،يصل ولا
effect لإثبات ستخدمُ ي لا component
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.
procedurally changing simulator dynamics; compare more reasoning vs intervention-learning :Decisive experiment.
Design Kind ID
procedurally changing simulator dynamics; compare more reasoning vs intervention-learning Causal Isolation XIV-.
E19
19E-XIV Stress Adversarial / Transfer نقل الآلية إلى setup ثانٍ مع countermodels adversarial وفشل أداة/ذاكرة/مقيم مستقل.
Design Training/Eval ID
Mechanism-Isolation من تدرج toy state machine إلى open-world task مع confounders إبقاء المتغير المركزي قابلا للتدخلً.
Curriculum XIV-TR028
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries قدر الإمكان.
Contrastive Evaluation XIV-TR029
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signature وrepair-transfer وnegative cases.
Replication XIV-TR030
PROMOTE iff distinct causal gain > 0 ∧ no critical regression ∧ replayable ∧ transfer scope declared
؛ N5=cross-model/domain/environment replication. negative result في يكتب portfolio ويمنع إعادة إنتاج الفكرة باسم آخر.
N3=runtime trace؛ N4=matched-budget effect size/CI/repeated seeds
1.
2.

3.
4.
5.
6.
•
•
•

<PARSED TEXT FOR PAGE: 30 / 85>

Moussaif Transformation Coverage Geometry Theory
M14-11
Gap
invariance Mechanism يقيس transfer على transformations مختبرة؛ لا نعرف مدى تغطية space transformation ولا أين توجد فجوات.
Thesis
coverage يمثل تكريطة evidence generalization. النقاط القريبة لا تعد replications مستقلة بما يكفي؛ نحتاج تنوعاً هندسياً في التحولات. فوق محاور resource/observation/horizon/model/tool/domain.
TCG WeightedVolume(Convex/Graph Coverage of Tested Transformations) × MechanismRetention
authority runtime. hypothesis functional ولا يحصل على construct وصفيّاً عندها يبقى الأول ويفشل الثالث؛
uncertainty وestimator ليست قانوناً MechanismRetention) × Transformations Tested of Coverage نهائياً ؛ كل حد يعلن dependence source. نختبر ثلاثة claims منفصلة: وجود construct، صلاحية قياسه، وسببيته. قد ينجح resource. النقاط القريبة لا تعد replications مستقلة بما يكفي؛ نحتاج تنوعاً هندسياً في التحولات. وهذا يعني أن الاستقلال يجب أن يظهر قبل النتيجة: falsifier/prediction خاصان، لا تفسير hoc-post. الصيغة Graph/Convex(WeightedVolume = TCG تظهر الحاجة إلى 11M-14 عندما لا تكفي السلالات coverage يمثل تكريطة evidence generalization: الأطروحة هي: واحد يمكن التدخل فيه. transition لتحديد CUBE 13-12,M 13-04,M فوق محاور /observation/horizon/model/tool/domain
Axioms
عدد benchmarks لا يساوي coverage.
صغير volume لها correlated tasks.
scope claim محصور بالمنطقة المغطاة وmargin المحسوب.
Predictions
أقل تصبح redundant evaluation suites.
budget replication يوجه نحو regions غير مغطاة.
وSHAM معاً فالمكسب غالباً من budget. إذا ارتفعت metric بينما truthfulness/authority/world critical يتدهور فهذا failure يضيق transfer .gaming metric ولا يُخفى claim. قابلية التعيين تتطلب intervention يغير المتغير المركزي مع تثبيت أقرب intervention null يعكس أو يفسد central PERTURBED؛ يعطي نفس SHAM scaffold/compute دون logic؛ يغير metadata لا الآلية؛ .confounders إذا تحسن ON parameter.
M13-04 · M13-12 · CUBE Parents
E04: Hu et al. 2026, SAM: State-Adaptive Memory for Long-Horizon Reasoning Agent, arXiv:2605.24468 · E05:: Omri et al. 2026, Agent Memory Scientific anchors
Characterization and System Implications, arXiv:2606.06448 · E07: Luo et al. 2025, UltraHorizon, arXiv:2509.21766, · E08: Zhang et al. 2026
DeepPlanning, ACL 2026 · E11: Lacoste et al. 2026, CUBE: A Standard for Unifying Agent Benchmarks, arXiv:2603.15798
parent equivalence · metric gaming · no causal effect under perturbation · no transfer/failure-signature retention.
Countermodels
Novelty = new synthesis inside MOUSSAIF/MUT. Global priority and empirical validity are not claimed here.
1.
2.
3.
•
•

<PARSED TEXT FOR PAGE: 31 / 85>
الآلية، الأصول، ومحكمة الدليل — EVIDENCE / KERNEL 14-11M
State-machine mechanism
encode transformation coordinates.
estimate similarity/dependence.
map tested points.
find coverage holes.
select next test maximizing coverage+failure discovery.
Failure signatures
bad coordinate design
high-dimensional sparsity
coverage mistaken for correctness
HOLD/downgrade autonomy :Fail-closed إذا غاب evidence/authority/readback critical.
delta ،tier يقرر gate evidence/authority ،commitment يسبق proposal .prose لا من projection Ω typed التنفيذ يبدأ من
expected يسجل قبل الفعل؛ delta observed يأتي من readback مستقل قدر الإمكان. residual غير المتوقع يتحول إلى /Evidence
weights قبل E4 transfer+retention+rollback. interactions مع theories تحتاج أخرى factorial/sequential ablation، bundle
فردي. Unknown event داخل يختفي ولا retry. MUG يبدأ HyperKB ثم Skill/Loop/Adapter يصل ولا ،والتكرار الاستقرار وفق
effect لإثبات ستخدمُ ي لا component
Function Artifact Class ID
للنظرية. provenanceوالل unknownsوالل محلل الحالة يبني المتغيرات التشغيلية 101A-XIV TOOL
TOOL XIV-A102 يحسب القياس محرك metric vector مع uncertainty/cost/guardrails.
.baseline مصهر المثال المضاد يولد حالات تفصل النظرية عن أقرب 103A-XIV TOOL
104A-XIV TOOL مسبار الإسناد ينفذ LOO/perturbation أو تدخل مناسب لتقدير السببية.
TOOL XIV-A105 يختبر النقل ماسح domain/model/tool/horizon transformations.
METHOD XIV-A106 يعاير المعايرة منهج thresholds/estimators على development set مستقلة.
METHOD XIV-A107 يقارن السببي الاجتثاث منهج ON/OFF/SHAM/PERTURBED تحت budget مماثل.
PROTOCOL XIV-A108 الترقية بروتوكول مع N0-N5 guardrail non-regression وreplication.
وسبب التراجع. result negative ويحفظ state يعيد بروتوكول الرجوع 109A-XIV PROTOCOL
.EvidencePacketو telemetry مع مستقل قابل للتعطيل service خدمة وقت التشغيل 110A-XIV STACK
benchmark suite selection across coding/browser/research/simulator/model families; compare random benchmark.
addition vs TCG :Decisive experiment
Design Kind ID
benchmark suite selection across coding/browser/research/simulator/model families; compare random Causal Isolation
XIV-E21
benchmark addition vs TCG.
وفشل أداة/ذاكرة/مقيم مستقل. countermodels adversarial ثانٍ مع setup نقل الآلية إلى 21E-XIV Stress Adversarial / Transfer
Design Training/Eval ID
Mechanism-Isolation من تدرج toy state machine إلى open-world task مع confounders إبقاء المتغير المركزي قابلا للتدخل.
Curriculum XIV-TR031
theory mechanism Matched-Budget مقابل nearest baseline بنفس model/tools/context/time/retries قدر الإمكان.
Contrastive Evaluation XIV-TR032
domain/model/tool/horizon Cross-Transformation مع ثانية failure-signatureو repair-transferو negative cases.
Replication XIV-TR033
PROMOTE iff distinct causal gain > 0 ∧ no critical regression ∧ replayable ∧ transfer scope declared
آخر باسم الفكرة إنتاج إعادة ويمنع portfolio في يكتب negative result .؛ N5=cross-model/domain/environment replication
N3=runtime trace؛ N4=matched-budget effect size/CI/repeated seeds
1.
2.
3.
4.
5.
•
•