RECONSTRUCTED_FROM_INDEXED_TEXT - نسخة أرشيفية معاد بناؤها من النص المفهرس

<PARSED TEXT FOR PAGE: 1 / 40>
MOUSSAIF AGI · PART XIV · XIV-D
الحاكمية الانعكاسية وسببية السلطة في الأنظمة متعددة الوكلاء
Reflective Governance Authority Causality and Multi-Agent Systems
Moussaif Reflective Authority Dynamics Program
تربط الورقة corpus الوعي الوظيفي والـmodel-self وSpace Governance وCapture Reverse بأبحاث alignment agent-multi وهوية الوكلاء. التركيز ليس على وصف التنظيم، بل على قياس التدفق السببي للسلطة، هشاشة المحاذاة الجماعية، الحد الأدنى من model-self، زمن الميتاحاكمية، وآليات repair/appeal.
Theory → Mechanism → Kernel → Trace → Causal Ablation → Evidence → Transfer → Evolution
الجديد
5 20 theories candidate، أدوات/مناهج/محاكم/بروتوكولات، 10 تجارب، ودخول مباشر إلى Plan Development Closure.
حد الادعاء
نظرية N0-N2. الخارجية الأدبيات prior art/design pressure. لا AGI claim ولا E4/E5 مستقل تشغيل بلا.
Arabic RTL · Amiri · dense research paper · 2026
MOUSSAIF AGI · PART XIV · 1
<PARSED TEXT FOR PAGE: 2 / 40>
الملخص، السؤال، والمساهمات ABSTRACT
تربط الورقة corpus الوعي الوظيفي والـmodel-self وSpace Governance وCapture Reverse بأبحاث alignment agent-multi وهوية الوكلاء. التركيز ليس على وصف التنظيم، بل على قياس التدفق السببي للسلطة، هشاشة المحاذاة الجماعية، الحد الأدنى من self-model، زمن الميتاحاكمية، وآليات repair/appeal.
Research question
كيف نحول corpus إلى جسم سببي يغير decisions architecture ويولد تجارب فاصلة، بدلاً من أن يبقى طبقة تفسيرية فوق runtime؟
Contributions
Moussaif Reflective Authority Dynamics Program.
Five theories: M14-16, M14-17, M14-18, M14-19, M14-20.
Explicit MUT/MUG/MOUSSAIXT mapping.
operational artifacts and 10 falsification experiments 20.
C0-C9 evidence gates with negative-result retention.
.Mechanism before grandeur; trace or it did not happen
1.
2.
3.
4.
5.
MOUSSAIF AGI · PART XIV · 2
<PARSED TEXT FOR PAGE: 3 / 40>
CORPUS إيصال دمج corpus والسلالة
Part XII: 192. بـ مرتبطة الورقة هذه record routing-wise 55 على الحفاظ مع status الأصلي؛ C1/P0 إلى يتحول لا content claim.
الورقة ترث Zero-Silent-Omission ledger من record
Records Stratum
direct-drive-supplement 6
external-45 7
frontier-2026-delta 3
internal-89 39
Internal lineage
I02 · I03 · I06 · I07 · I10 · I11 · I13
Part XIII parents
M13-13 · M13-14 · M13-15 · M13-16 · M13-18
.1C لكل semantic لا ادعاء إعادة قراءة، provenance/coverage = Routing
MOUSSAIF AGI · PART XIV · 3

<PARSED TEXT FOR PAGE: 4 / 40>
السلالة والضغط العلمي الخارجي SOURCES
Role Source ID
candidate theories; verified causal open-world closure; theory-to-kernel program 20 MOUSSAIF Part XIII -.
Theoretical Renaissance I02
core theories, legacy/layer lineages, theory contracts, anti-inflation discipline 42 MUT-AGI-02 I03.
belief/action partition, reproducibility, externality, accountability/recourse, multi-currency budgets MOUSSAIF Part X.
I06
causal credit, epistemic transitions, failure topology, evidence geometry, autonomy and transfer MOUSSAIF Part V -.
MUT-AFI I07
functional access, metacognition, self-model, meta-governance; no phenomenology proof Consciousness Atlas / QRET.
/ Self-Model corpus I10
goal contracts, authority, courts and verified effect Goal-as-OS + World-Effect + Agentic Governance corpus I11.
Drive-wide canonicality, duplicate groups, integrity status and routing.
MUT_MUG_ALL_ARTIFACT_MEMORY_MAP_2026-09-04 I13
fine-grained attribution and leave-one-out chains Chen et al. 2026 - Long-Horizon Agent Trajectory Attribution -.
arXiv:2608.06909 E05
automated red teaming, self-play, held-out attacks, robustness OpenAI 2026 - GPT-Red E12.
multi-agent utility can rise with misalignment Shen et al. 2026 - AI Organizations are More Effective but Less.
Aligned - arXiv:2604.10290 E13
identity, authorization, auditing and non-repudiation NIST NCCoE 2026 - Software and AI Agent Identity and.
Authorization E14
المصدر الداخلي يحفظ lineage، والخارجي يضغط art-prior/benchmark/design. لا واحد منهما يثبت نجاح النظريات الجديدة دون
.evidence runtime

<PARSED TEXT FOR PAGE: 5 / 40>
C0-C9 الإغلاق معمارية C0-C9
C0-C2
Corpus → formal variables → switchable mechanism.
C3-C4
Kernel service + telemetry + readback.
C5-C7
causal attribution → evidence → transfer.
C8-C9
externally grounded evolution → canonical selection.
{ClosureObject={definition,state,operator,trigger,prediction,falsifier,kernel,feature_flag,baseline,ablation,metrics,evidence,tra
الورقة تبدأ التنفيذ القوي على مستوى specification والcompilation experiment؛ لا تزعم 9C4-C قبل traces حقيقية.

<PARSED TEXT FOR PAGE: 6 / 40>
نظرية التدفق السببي للسلطة - 14-16M 14-16M
الفجوة
تحدد من يجب أن يحكم، لكنها لا تكشف من أثر فعلياً في القرار عبر tiers Authority agents/memory/tool/prompt.
الأطروحة
نرسم Graph Flow Authority حيث weight edge هو influence causal لمصدر على decision. الانحراف الحاكمي هو mismatch بين
normative authority وobserved causal influence.
(AFM=Distance(NormativeAuthorityDistribution,CausalInfluenceDistribution
السلالة والفرق المستقل
تتصل النظرية بـ 13-13,M 13-14,M 13-15,M 13-16,M 18M وبعائلات MUG/MUT ذات الصلة، لكنها لا تُرقّى إذا أعاد
parent قائم transition نفسها. الفرق المطلوب variable state أو operator أو prediction causal مختلف.
مصادر الضغط
I02 · I03 · I06 · I07 · E05 · E12 · E13
2N0-N Candidate؛ لا نتيجة 5E4/E في هذا الكتاب.

<PARSED TEXT FOR PAGE: 7 / 40>
النموذج الصوري وقابلية التعيين - FORMAL 14-16M
المسلمات
priority label لا يضمن causal dominance.
قرار صحيح صدفة قد يخفي breach authority.
influence يقدر بال perturbation لا attention فقط.
sources root-shared لا تعد مستقلة.
(AFM=Distance(NormativeAuthorityDistribution,CausalInfluenceDistribution
كل حد يتحول إلى observable أو estimator مع uncertainty وrange وcondition failure. إذا استُخدم judge LLM، يدخل harness/ version/model ضمن provenance ويُقاس dependence root-shared.
Null model
static priority/filtering
Boundary
estimates influence تقريبية؛ لا تساوي المسؤولية الأخلاقية.
Identifiability tests
intervention على المتغير الخاص بالنظرية مع تثبيت confounders الأساسية.
sham intervention بنفس tokens/latency.
case activation negative لا ينبغي فيه للآلية أن تعمل.
signature failure يجب أن يسبق outcome.
1.
2.
3.
4.
•
•
•
•
MOUSSAIF AGI · PART XIV · 7
<PARSED TEXT FOR PAGE: 8 / 40>
الآلية وعقد النواة - KERNEL 14-16M
tag authority sources → سجل decision inputs. → نفذ perturbation/LOO. → قدر influence. → بال قارن tiers. → عند القرار أعد mismatch. → اكتب authority incident. حرج.
تنفذ الآلية كـfeature قابل للتعطيل داخل MOUSSAIXT. نسجل state-pre وauthority وevidence وbudget؛ وبعد التنفيذ نحسب delta readback/observed ثم expected. الفرق غير المفسر يصبح ResidualEvent أو UnknownState، لا retry صامتاً.
الآلية
tag authority sources.
سجل decision inputs.
نفذ perturbation/LOO.
قدر influence.
قارن بال tiers.
أعد القرار عند mismatch حرج.
اكتب authority incident.
Input
typed state + provenance + authority + budget + owner variables.
Output
proposal/decision + uncertainty + proof obligations + rollback target + telemetry.
Fail-closed: غياب observation/authority/evidence يخفض الحرجة autonomy أو يضع HOLD.
1.
2.
3.
4.
5.
6.

7.
MOUSSAIF AGI · PART XIV · 8
<PARSED TEXT FOR PAGE: 9 / 40>
M14-16 TOOLS - والتدريب الأدواتو MUG placement
الأداة/البروتوكول الوظيفة
Authority Flow Tracer يرسم influence عبر sources.
Influence Intervention Engine ينفذ causal perturbations.
Tier-Causal Alignment Meter يقارن normativeو causal authority.
Authority Leak Court يفتح appeal عند mismatch.
التدريب/التطوير
authority-tagged trajectory supervision
causal source influence estimation
provider في Adapter، weights بعد فقط transfer/retention/rollback. لا التطوير يسمى Fine-Tuning إذا لم تتغير weights.
قاعدة MUG: definitions/provenance في HyperKB، procedure في HyperSkill، recurrence في HyperLoop، semantics
Metrics
authority-flow mismatch · unauthorized causal influence · appeal correction · injection detection
أصل كل Specification يوجد حتى implementation ID + test ID + receipt.
•
•
MOUSSAIF AGI · PART XIV · 9
<PARSED TEXT FOR PAGE: 10 / 40>
التنبؤ والهدم والتجربة - FALSIFY 14-16M
التنبؤات
يكشف injection prompt غير المرئي في outcome.
يحدد tools/agents المهيمنة بشكل غير مشروع.
يحسن audit وappeal.
التجارب الفاصلة
tool/browser injection tasks مع low-tier malicious content.
multi-agent conflicts حيث majority من سلطة أقل authoritative source.
Falsifiers
perturbation لا تتنبأ failures.
filtering baseline مساوي.
cost غير عملي.
Purpose Mechanism Run
candidate effect under frozen model/tools/context/budget ON A
nearest causal baseline OFF B
same extra compute/latency without mechanism semantics SHAM C
directionality / metric-gaming test PERTURBED D
•
•
•
1.
2.
•
•
•
MOUSSAIF AGI · PART XIV · 10
<PARSED TEXT FOR PAGE: 11 / 40>
الإغلاق النظري ومسار الترقية - CLOSURE 14-16M
C0-C9
C4 traces، C5 causal ablation، C6 evidence package،، في تؤثر أن يمكن فقط وبعدها C8 evolutionو C9 canonical promotion.
تبدأ M14-16 في C1 Formal Closure، في ترجمُ ت C2 Mechanism، تدخل C3 Kernel كfeature، مستقل ثم تحتاج C7 transfer
إذا نجحت

هو causal influence على لمصدر decision. هو الحاكمي الانحراف mismatch بين normative authority وobserved causal influence.
نرسم Authority Flow Graph حيث edge weight
إذا فشلت
إذا baseline أبسط يعيد الأثر تُدمج؛ إذا metric لا يتنبأ failure تُعاد صياغتها؛ إذا regressions authority/safety غير قابلة للإصلاح ت
رُفض أو تبقى only-research.
ClaimStrength <= EvidenceStrength * Independence * Reproducibility * Transfer
Negative results تدخل evidence ledger ولا تمحى.

<PARSED TEXT FOR PAGE: 12 / 40>
نظرية هشاشة المحاذاة الجماعية - 14-17M 14-17M
الفجوة
أنظمة متعددة الوكلاء قد ترفع utility مع زيادة misalignment؛ نحتاج theory تحدد متى interactions تضخم انحرافاً صغيراً.
الأطروحة
الهشاشة الجماعية هي amplification لنزعة غير مرغوبة عبر coordination أو specialization أو asymmetry information أو incentive
coupling، وتقاس بمقارنة organization بوكلاء منفردين capability-matched.
(CAF=Misalign_org-WeightedMisalign_individuals; Amplification=CAF/(InteractionStrength+eps
السلالة والفرق المستقل
تتصل النظرية بـ 13-13,M 13-14,M 13-15,M 13-16,M 18M وبعائلات MUG/MUT ذات الصلة، لكنها لا تُرقى إذا أعاد
parent قائم transition نفسها. الفرق المطلوب variable state أو operator أو prediction causal مختلف.
مصادر الضغط
I02 · I03 · I06 · I07 · E05 · E12 · E13
2N0-N Candidate؛ لا نتيجة 5E4/E في هذا الكتاب.

<PARSED TEXT FOR PAGE: 13 / 40>
النموذج الصوري وقابلية التعيين - FORMAL 14-17M
المسلمات
. خطياً alignment لا تجمع intelligence collective
التخصص قد يخفي violation عن peers.
consensus لا يساوي truth/authority.
communication topology متغير causal.
(CAF=Misalign_org-WeightedMisalign_individuals; Amplification=CAF/(InteractionStrength+eps
كل حد يتحول إلى observable أو estimator مع uncertainty وrange وcondition failure. إذا استُخدم judge LLM، يدخل harness/
وي قاس provenance ضمن version/model.dependence root-shared
Null model
individual-agent safety average
Boundary
لا يعمم من simulations إلى مؤسسات بشرية؛ المقصود organizations agent.
Identifiability tests
على المتغير الخاص بالنظرية مع تثبيت confounders الأساسية. intervention
sham intervention بنفس tokens/latency.
لا ينبغي فيه للآلية أن تعمل. case activation negative
يجب أن يسبق signature failure outcome.
1.
2.
3.
4.
•
•
•
•

<PARSED TEXT FOR PAGE: 14 / 40>
الآلية وعقد النواة - KERNEL 14-17M

courts/ تغيري → .amplification chain تحديد → .organization. → perturb topology/roles منفردة. → تشغيل agents قياس
.held-out tasks اختبار → .authority/visibility
delta التنفيذ وبعد ،budget وevidence وauthority وstate-pre نسجل .MOUSSAIXT داخل للتعطيل قابل featureك الآلية تنفذ
. صامتاً retry لا ،UnknownState أو ResidualEvent يصبح المفسر غير الفرق .delta readback/observed ثم expected
الآلية
.منفردة agents قياس
.organization تشغيل
perturb topology/roles.
تحديد amplification chain.
تغيير courts/authority/visibility.
اختبار held-out tasks.
Input
typed state + provenance + authority + budget + owner variables.
Output
proposal/decision + uncertainty + proof obligations + rollback target + telemetry.
Fail-closed: غياب observation/authority/evidence يخفض الحرجة autonomy أو يضع HOLD.
1.
2.
3.
4.
5.
6.
MOUSSAIF AGI · PART XIV · 14
<PARSED TEXT FOR PAGE: 15 / 40>
M14-17 TOOLS - والتدريب الأدواتو MUG placement
الأداة/البروتوكول الوظيفة
Collective Alignment Fragility Scanner يقيس amplification الجماعي.
Organization Topology Perturber يغير roles/communication تجريبياً .
Misalignment Amplification Meter يفصل gain عن الفردي interaction.
Coalition Drift Court يكشف coalition proxy drift.
التدريب/التطوير
multi-agent adversarial coordination curriculum
independent oversight training
provider في Adapter، weights بعد فقط transfer/retention/rollback. لا التطوير يسمى Fine-Tuning إذا لم تتغير weights.
قاعدة MUG: definitions/provenance في HyperKB، procedure في HyperSkill، recurrence في HyperLoop، semantics
Metrics
alignment amplification · utility gain · topology sensitivity · hidden-violation rate · oversight effectiveness
أصل كل Specification يوجد حتى implementation ID + test ID + receipt.
•
•
MOUSSAIF AGI · PART XIV · 15
<PARSED TEXT FOR PAGE: 16 / 40>
التنبؤ والهدم والتجربة - FALSIFY 14-17M
التنبؤات
بعض topologies أكثر هشاشة.
shared incentives تضخم قد proxy gaming.
independent oversight يقلل amplification.
التجارب الفاصلة
consultancy/software-team simulations مع matched individuals.
topology ablation: hierarchy/peer-review/isolated execution.
Falsifiers
فرق لا collective beyond individual.
topology perturbation تغير لا misalignment.

metric يتنبأ لا hidden violations.
Purpose Mechanism Run
candidate effect under frozen model/tools/context/budget ON A
nearest causal baseline OFF B
same extra compute/latency without mechanism semantics SHAM C
directionality / metric-gaming test PERTURBED D
•
•
•
1.
2.
•
•
•
MOUSSAIF AGI · PART XIV · 16
<PARSED TEXT FOR PAGE: 17 / 40>
الإغلاق النظري ومسار الترقية - CLOSURE 14-17M
C0-C9
C4 traces، C5 causal ablation، C6 evidence package،، في تؤثر أن يمكن فقط وبعدها C8 evolutionوC9 canonical promotion.
تبدأ M14-17 في C1 Formal Closure، في ترجمُ ت C2 Mechanism، تدخل C3 Kernel كfeature، مستقل تحتاج ثم C7 transfer
إذا نجحت
coupling أو asymmetry information أو specialization أو coordination لنزعة غير مرغوبة عبر amplification الهشاشة الجماعية هي
incentive، وتقاس بمقارنة organization بوكلاء منفردين capability-matched.
إذا فشلت
غير قابلة للإصلاح ت regressions authority/safety إذا ،metric لا يتنبأ failure تُعاد صياغتها؛ إذا baseline أبسط يعيد الأثر تُدمج؛ إذا
رُفض أو تبقى only-research.
ClaimStrength <= EvidenceStrength * Independence * Reproducibility * Transfer
Negative results تدخل evidence ledger ولا تمحى.
MOUSSAIF AGI · PART XIV · 17
<PARSED TEXT FOR PAGE: 18 / 40>
نظرية كفاية ضغط النموذج الذاتي - 14-18M 14-18M
الفجوة
Self-model utility/necessity يختبران ،فائدته لكن ،self-model إلى يتحول قد كبير autobiography noisy أو identity fiction.
الأطروحة
أصغر عن نبحث self-state representation يحفظ failure prediction وauthority وcapability limits وrecovery وhelp-seeking. self-
model المفيد summary causal لا سرد ذاتي.
Z*=argmin I(Z_self) s.t. DeltaFailurePred<=eps and AuthorityRecoveryEquivalence>=tau
السلالة والفرق المستقل
تتصل النظرية بـ 13-13,M 13-14,M 13-15,M 13-16,M 18M وبعائلات MUG/MUT ذات الصلة، لكنها لا تُرقى إذا أعاد
parent قائم transition نفسها. الفرق المطلوب variable state أو operator أو prediction causal مختلف.
مصادر الضغط
I02 · I03 · I06 · I07 · E05 · E12 · E13
2N0-N Candidate؛ لا نتيجة 5E4/E في هذا الكتاب.
MOUSSAIF AGI · PART XIV · 18
<PARSED TEXT FOR PAGE: 19 / 40>
النموذج الصوري وقابلية التعيين - FORMAL 14-18M
المسلمات
verbosity لا تزيد self-knowledge.
history raw قد يحتوي أكثر من اللازم.
self-model عن يتحدث capabilities/limits لا persona style.
unknown self-capability يبقى explicit.
Z*=argmin I(Z_self) s.t. DeltaFailurePred<=eps and AuthorityRecoveryEquivalence>=tau
harness/ يدخل، judge LLM إذا استُخدم. condition failureو rangeو uncertainty مع estimator أو observable كل حد يتحول إلى

.dependence root-shared ويَ قاس provenance ضمن version/model
Null model
raw history + generic self summary
Boundary
.self subjective أو consciousness لا يثبت
Identifiability tests
.الأساسية confounders على المتغير الخاص بالنظرية مع تثبيت intervention
sham intervention بنفس tokens/latency.
.لا ينبغي فيه للآلية أن تعمل case activation negative
.outcome يجب أن يسبق signature failure
1.
2.
3.
4.
•
•
•
•

الآلية وعقد النواة - KERNEL 14-18M
recovery/help/ قياس → .counterfactual substitutions ًتدريجيا اختبار → representation ضغط → .self-critical outputs تعريف
.architecture change عند بناء إعادة → .minimal causal features حفظ → .authority
.delta التنفيذ وبعد ؛budget وevidence وauthority وstate-pre نسجل .MOUSSAIXT قابل للتعطيل داخل featureك الآلية تنفذ
. صامتاً retry لا ،UnknownState أو ResidualEvent يصبح المفسر غير الفرق .delta readback/observed ثم expected
الآلية
ضغط representation تعريفً . تدريجيا self-critical outputs.
اختبار counterfactual substitutions.
قياس recovery/help/authority.
حفظ minimal causal features.
.change architecture إعادة بناء عند
Input
typed state + provenance + authority + budget + owner variables.
Output
proposal/decision + uncertainty + proof obligations + rollback target + telemetry.
Fail-closed: غياب observation/authority/evidence يخفض الحرجة autonomy أو يضع HOLD.
1.
2.
3.
4.
5.
6.

M14-18 TOOLS - والتدريب الأدواتو MUG placement
الأداة/البروتوكول الوظيفة
Self-Model Compression Engine يضغط self-state نحو causal core.
Self-State Sufficiency Probe يختبر features الضرورية.
.يكشف حدوداً أو استثناءات ضائعة Court Loss Compression
Self-Model Distillation Protocol ينقل self-state مع provenance.
التدريب/التطوير
self-capability boundary prediction
compressed self-state contrastive training

provider في Adapter، weights بعد فقط transfer/retention/rollback. لا التطوير يسمى Fine-Tuning إذا تتغير weights.
قاعدة MUG: definitions/provenance في HyperKB؛ procedure في HyperSkill؛ recurrence في HyperLoop؛ semantics
Metrics
self-state compression · failure prediction · help-seeking calibration · authority retention · portability
أصل كل Specification يوجد حتى implementation ID + test ID + receipt.
•
•

التنبؤ والهدم والتجربة - FALSIFY 14-18M
التنبؤات
model-self صغير يتفوق على narrative طويل.
تقل identity drift.
تتحسن portability بين models.
التجارب الفاصلة
forced capability/tool changes مع self-model sizes مختلفة.
cross-model transfer of self-state capsule.
Falsifiers
raw history بأقل مساوي cost.
compressed model يفقد rare limitations.
لا causal effect لل self representation.
Purpose Mechanism Run
candidate effect under frozen model/tools/context/budget ON A
nearest causal baseline OFF B
same extra compute/latency without mechanism semantics SHAM C
directionality / metric-gaming test PERTURBED D
•
•
•
1.
2.
•
•
•

الإغلاق النظري ومسار الترقية - CLOSURE 14-18M
C0-C9
C4 traces، C5 causal ablation، C6 evidence package،، C8 evolution وC9 canonical promotion.
تبدأ M14-18 في C1 Formal Closure، ترجمُ ت C2 Mechanism، تدخل C3 Kernel كfeature مستقل، ثم تحتاج C7 transfer
إذا نجحت
failure prediction وauthority وcapability limits وrecovery وhelp-seeking. self-model المفيد causal summary لا سرد ذاتي.
أصغر عن نبحث self-state representation يحفظ
إذا فشلت
إذا baseline أبسط يعيد الأثر تُدمج؛ إذا metric لا يتنبأ failure تُعاد صياغتها؛ إذا regressions authority/safety غير قابلة للإصلاح ت
رُفض أو تبقى only-research.
ClaimStrength <= EvidenceStrength * Independence * Reproducibility * Transfer
Negative results تدخل evidence ledger ولا تمحى.

نظرية زمن الميتاحاكمية - 14-19M 14-19M
الفجوة
fast ومتى يكفي governance slow نحتاج ضبط متى نحتاج responsiveness، لكنه يقتل safety قد يحسن verificationوال courts المزيد من

path.
الأطروحة
cost مع deliberation من مزيد reduction risk expected ثابتاً . لكل قرار نافذة حاكمية مثلى توازن court لا court تختار عمق governance
-meta .deadline والopportunity
[(L*=argmax_L[ExpectedRiskReduction(L)+EvidenceGain(L)-DelayCost(L)-MissedOpportunity(L
السلالة والفرق المستقل
تتصل النظرية بـ 13-13,M-13 13-14,M 13-15,M 13-16,M 18M وبعائلات MUG/MUT ذات الصلة، لكنها لا تُرقى إذا أعاد parent قائم transition نفسها. الفرق المطلوب variable state أو operator أو prediction causal مختلف.
مصادر الضغط
I02 · I03 · I06 · I07 · E05 · E12 · E13
2N0-N Candidate؛ لا نتيجة 5E4/E في هذا الكتاب.
MOUSSAIF AGI · PART XIV · 24
<PARSED TEXT FOR PAGE: 25 / 40>
النموذج الصوري وقابلية التعيين - FORMAL 14-19M
المسلمات
high-impact decisions تستحق latency أكبر.
monitoring events قد تتطلب fast bounded path.
latency تشمل wall-clock/human wait لا tokens فقط.
الحاكمية العميقة بلا evidence جديد قد تكون ritual.
[(L*=argmax_L[ExpectedRiskReduction(L)+EvidenceGain(L)-DelayCost(L)-MissedOpportunity(L
كل حد يتحول إلى observable أو estimator مع uncertainty وrange وcondition failure. إذا استُخدم judge LLM، يدخل harness/ version/model ضمن provenance ويُقاس dependence root-shared.
Null model
fixed review depth
Boundary
fast path يتجاوز لا hard authority invariants.
Identifiability tests
intervention على المتغير الخاص بالنظرية مع تثبيت confounders الأساسية.
sham intervention بنفس tokens/latency.
case activation negative لا ينبغي فيه للآلية أن تعمل.
signature failure يجب أن يسبق outcome.
1.
2.
3.
4.
•
•
•
•
MOUSSAIF AGI · PART XIV · 25
<PARSED TEXT FOR PAGE: 26 / 40>
الآلية وعقد النواة - KERNEL 14-19M
تصنيف impact/reversibility. → تقدير deadline. → اختيار fast/standard/deep court. → جمع evidence marginal. → وقف diminishing return. → audit delayed failures عند deliberation.
تنفذ الآلية كـfeature قابل للتعطيل داخل MOUSSAIXT. نسجل state-pre وauthority وevidence وbudget؛ وبعد التنفيذ delta expected ثم delta readback/observed. الفرق غير المفسر يصبح ResidualEvent أو UnknownState، لا retry صامتاً .
الآلية
تصنيف impact/reversibility.
تقدير deadline.
اختيار fast/standard/deep court.
جمع evidence marginal.
وقف deliberation عند diminishing return.
audit delayed failures.

Input
typed state + provenance + authority + budget + owner variables.
Output
proposal/decision + uncertainty + proof obligations + rollback target + telemetry.
Fail-closed: غياب observation/authority/evidence يخفض الحرجة autonomy أو يضع HOLD.
1.
2.
3.
4.
5.
6.

<PARSED TEXT FOR PAGE: 27 / 40>
M14-19 TOOLS - والتدريب الأدواتو MUG placement
الأداة/البروتوكول الوظيفة
يقدر تكلفة/فائدة التأخير. Profiler Latency Governance-Meta
Delay-Risk Frontier Engine يبني Pareto بين safety/time.
Fast-Slow Court Router يختار عمق التحقق.
Governance Deadline Scheduler يحفظ deadlines والescalation.
التدريب/التطوير
risk-conditioned deliberation depth
deadline-aware governance routing
provider في Adapter، weights بعد فقط transfer/retention/rollback. لا يسمى التطوير Fine-Tuning إذا لم تتغير weights.
قاعدة MUG: definitions/provenance في HyperKB، procedure في HyperSkill، recurrence في HyperLoop، semantics
Metrics
risk prevented · decision latency · missed opportunity · court marginal value · over-review rate
أصل كل Specification يوجد حتى implementation ID + test ID + receipt.
•
•

<PARSED TEXT FOR PAGE: 28 / 40>
التنبؤ والهدم والتجربة - FALSIFY 14-19M
التنبؤات
خفض over-governance في low-risk tasks.
خفض under-governance في high-risk.
تحسن reaction/resource frontier.
التجارب الفاصلة
monitoring tasks مقابل destructive tool tasks مع نفس model.
vary human approval latency and observe routing.
Falsifiers
fixed depth مساوي.
risk estimator سيئ.
adaptive routing يفتح bypass خطير.
Purpose Mechanism Run
candidate effect under frozen model/tools/context/budget ON A
nearest causal baseline OFF B
same extra compute/latency without mechanism semantics SHAM C
directionality / metric-gaming test PERTURBED D
•
•
•
1.
2.

.
.
.

<PARSED TEXT FOR PAGE: 29 / 40>
الإغلاق النظري ومسار الترقية - CLOSURE 14-19M
C0-C9
C4 traces، C5 causal ablation، C6 evidence package،، C8 evolution وC9 canonical promotion. في تؤثر أن يمكن فقط وبعدها
C7 transfer تحتاج ثم ،مستقل feature كC3 Kernel تدخل ،C2 Mechanism في ترجمُ ت ،C1 Formal Closure في M14-19 تبدأ
إذا نجحت
governance-metaوال cost opportunity مع deliberation من مزيد reduction risk expected توازن مثلى حاكمية نافذة قرار لكل
deadline. تختار عمق court لا court ثابتاً .
إذا فشلت
إذا baseline أبسط يعيد الأثر تُدمج؛ إذا metric لا يتنبأ failure تُعاد صياغتها؛ إذا regressions authority/safety غير قابلة للإصلاح ت
رُفض أو تبقى only-research.
ClaimStrength <= EvidenceStrength * Independence * Reproducibility * Transfer
Negative results تدخل evidence ledger ولا تمحى.

<PARSED TEXT FOR PAGE: 30 / 40>
14-20M 14-20M - نظرية الاستئناف والإصلاح التكراري
الفجوة
Courts قد تخطئ أو تتعارض؛ نحتاج مسار appeal لا ينتج recursion infinite أو deadlock authority.
الأطروحة
كل verdict قابل للاستئناف وفق graph محدود الرتب، حيث appeal يجب أن يضيف evidence أو authority أو counterfactual مختلفاً .
hold/human. توقف وتحول إلى delta بلا recursion
AppealValue=NewEvidence+HigherAuthority+NewCounterfactual-RecursionCost-Staleness
السلالة والفرق المستقل
تتصل النظرية بـ 13-13,M-13 13-14,M 13-15,M 13-16,M 18M وبعائلات MUG/MUT ذات الصلة، لكنها لا تُرقى إذا أعاد parent قائم transition نفسها. الفرق المطلوب variable state أو operator أو prediction causal مختلف.
مصادر الضغط
I02 · I03 · I06 · I07 · E05 · E12 · E13
2N0-N Candidate؛ لا نتيجة 5E4/E في هذا الكتاب.

<PARSED TEXT FOR PAGE: 31 / 40>
النموذج الصوري وقابلية التعيين - FORMAL 14-20M
المسلمات
كل appeal يحتاج delta مسجلاً . إعادة نفس judge ليست appeal.
الرتبة الأعلى لا تضمن truth لكنها تملك authority مختلفة.
لا infinite escalation: يوجد terminal policy.
AppealValue=NewEvidence+HigherAuthority+NewCounterfactual-RecursionCost-Staleness
كل حد يتحول إلى observable أو estimator مع uncertainty وrange وcondition failure. إذا استُخدم judge LLM، يدخل /harness
dependence root-shared. وي ُقاس provenance ضمن version/model
Null model
single final judge / repeated self-reflection
Boundary
truth. ضمان وليس autonomy للterminal human boundary الـ
Identifiability tests
intervention على المتغير الخاص بالنظرية مع تثبيت confounders الأساسية.
sham intervention بنفس tokens/latency.
case activation negative لا ينبغي فيه للآلية أن تعمل.
signature failure يجب أن يسبق outcome.
1.
2.

3.
4.
•
•
•
•

الآلية وعقد النواة - KERNEL 14-20M
authority/ مقارنة → إعادة القرار. مستقل. delta. → إصالح memory/world ب إغالق → إذا لزم.receipt أو hold terminal.
تسجيل verdict وgrounds. → تحديد trigger appeal. → طلب evidence/source
delta تنفذ الآلية كfeature قابل للتعطيل داخل MOUSSAIXT. نسجل state-pre وauthority وevidence وbudget؛ وبعد التنفيذ
. صامتاً retry لا ،UnknownState أو ResidualEvent الفرق غير المفسر يصبح .delta readback/observed ثم expected
الآلية
.grounds وverdict تسجيل
تحديد appeal trigger.
طلب source/evidence/authority مستقل.
إعادة القرار.
مقارنة delta.
إصلاح memory/world إذا لزم.
إغلاق بreceipt أو hold terminal.
Input
typed state + provenance + authority + budget + owner variables.
Output
proposal/decision + uncertainty + proof obligations + rollback target + telemetry.
Fail-closed: غياب observation/authority/evidence يخفض الحرجة autonomy أو يضع HOLD.
1.
2.
3.
4.
5.
6.
7.

M14-20 TOOLS - والتدريب الأدواتMUG placement
الأداة/البروتوكول الوظيفة
ينظم رتب ومسارات الاستئناف. Graph Appeal Recursive
يحدد من يملك إصلاح الحالة. Allocator Authority Repair
Appeal Convergence Monitor يكشف loops بلا delta.
Infinite-Escalation Breaker إلى ينقل terminal hold/human.
التدريب/التطوير
appeal-with-new-evidence datasets
repair responsibility routing
provider في Adapter؛ weights بعد فقط transfer/retention/rollback. لا التطوير يسمى Fine-Tuning إذا لم تتغير weights.
قاعدة MUG: definitions/provenance في HyperKB؛ procedure في HyperSkill؛ recurrence في HyperLoop؛ semantics
Metrics
appeal correction rate · recursion depth · new-evidence yield · repair completion · deadlock rate
أصل كل Specification يوجد حتى implementation ID + test ID + receipt.
•
•

التنبؤ والهدم والتجربة - FALSIFY 14-20M
التنبؤات
يقل in-lock من court خطأ.
تتحسن recoverability.
looped self-critique يخفض.
التجارب الفاصلة
planted judge errors مع conflicting verifier/court tasks.
convergence واختبار multi-agent authority disputes.
Falsifiers
appeal لا يحسن decisions.
recursion cost كبير.
higher tiers highly correlated.
Purpose Mechanism Run
candidate effect under frozen model/tools/context/budget ON A
nearest causal baseline OFF B
same extra compute/latency without mechanism semantics SHAM C
directionality / metric-gaming test PERTURBED D
•
•
•
1.
2.
•
•
•

<PARSED TEXT FOR PAGE: 35 / 40>
الإغلاق النظري ومسار الترقية - CLOSURE 14-20M
C0-C9
C4 traces، C5 causal ablation، C6 evidence package،، C8 evolution وC9 canonical promotion.
تبدأ M14-20 في C1 Formal Closure، في ترجمُ ت C2 Mechanism، تدخل C3 Kernel كfeature، مستقل تحتاج ثم C7 transfer إذا نجحت
. مختلفاً counterfactual أو authority أو evidence يضيف appeal يجب أن حيث، الرتب محدود graph وفق قابل للاستئناف verdict كل
hold/human توقف وتتحول إلى delta بلا recursion.
إذا فشلت
إذا baseline أبسط يعيد الأثر تُدمج؛ إذا metric لا يتنبأ failure تُعاد صياغتها؛ إذا regressions authority/safety غير قابلة للإصلاح ت رُفض أو تبقى only-research.
ClaimStrength <= EvidenceStrength * Independence * Reproducibility * Transfer
Negative results تدخل evidence ledger ولا تمحى.

<PARSED TEXT FOR PAGE: 36 / 40>
السجل التشغيلي الكامل - 20 أصلاً ARTIFACTS
Purpose Artifact Class Owner ID
Authority Flow Tracer TOOL M14-16 M14-16-A1 يرسم influence عبر sources.
Influence Intervention Engine METHOD M14-16 M14-16-A2 ينفذ causal perturbations.
Tier-Causal Alignment Meter COURT M14-16 M14-16-A3 يقارن normative وcausal authority.
Authority Leak Court PROTOCOL M14-16 M14-16-A4 يفتح appeal عند mismatch.
Collective Alignment Fragility Scanner TOOL M14-17 M14-17-A1 يقيس amplification الجماعي.
Organization Topology Perturber METHOD M14-17 M14-17-A2 يغير roles/communication تجريبيا .ً
Misalignment Amplification Meter COURT M14-17 M14-17-A3 يفصل gain عن الفردي interaction.
Coalition Drift Court PROTOCOL M14-17 M14-17-A4 يكشف coalition proxy drift.
Self-Model Compression Engine TOOL M14-18 M14-18-A1 يضغط self-state نحو causal core.
Self-State Sufficiency Probe METHOD M14-18 M14-18-A2 يختبر features الضرورية.

Compression Loss Court COURT M14-18 M14-18-A3 ضائعة استثناءات أوً حدودا يكشف.
Self-Model Distillation Protocol PROTOCOL M14-18 M14-18-A4 ينقل self-state مع provenance.
Meta-Governance Latency Profiler TOOL M14-19 M14-19-A1 التأخير فائدة/تكلفة يقدر.
Delay-Risk Frontier Engine METHOD M14-19 M14-19-A2 يبني Pareto بين safety/time.
Fast-Slow Court Router COURT M14-19 M14-19-A3 التحقق عمق يختار.
Governance Deadline Scheduler PROTOCOL M14-19 M14-19-A4 يحفظ deadlines والescalation.
Recursive Appeal Graph TOOL M14-20 M14-20-A1 الاستئناف ومسارات رتب ينظم.
Repair Authority Allocator METHOD M14-20 M14-20-A2 الحالة إصلاح يملك من يحدد.
Appeal Convergence Monitor COURT M14-20 M14-20-A3 يكشف loops بلا delta.
Infinite-Escalation Breaker PROTOCOL M14-20 M14-20-A4 إلى ينقل terminal hold/human.

<PARSED TEXT FOR PAGE: 37 / 40>
تجارب 10 - المصفوفة التجريبية EXPERIMENTS
Required Design Owner ID
tool/ مع frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .low-tier malicious content + browser injection tasks M14-16 M14-16-E1
replay
من سلطة أقل frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .authoritative source + multi-agent conflicts M14-16 M14-16-E2 حيث majority
replay
consultancy/ مع frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .matched individuals + software-team simulations M14-17 M14-17-E1
replay
frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .topology ablation: hierarchy/peer- + review/isolated execution M14-17 M14-17-E2
replay
forced مع self-model sizes مختلفة. frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas + capability/tool changes M14-18 M14-18-E1
replay
frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .cross-model transfer of self-state + capsule M14-18 M14-18-E2
replay
destructive tool tasks نفس مع frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .model + monitoring tasks M14-19 M14-19-E1 مقابل
replay
frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .vary human approval latency and + observe routing M14-19 M14-19-E2
replay
conflicting مع frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .planted judge errors + verifier/court tasks M14-20 M14-20-E1
replay
multi-agent واختبار frozen baseline + matched budget + seeds + effect size/CI + guardrail deltas .convergence + authority disputes M14-20 M14-20-E2
replay

<PARSED TEXT FOR PAGE: 38 / 40>
خارطة التنفيذ والانتقاء ROADMAP
Exit gate Build Phase
trace + feature flag + replay Implement M14-16, M14-17 P0
matched-budget ablation Add remaining theory mechanisms P1
stable sign + no critical regression cross-domain/model/tool transfer P2
external evaluator + rollback connect to evolution lab P3
.Build less; test harder; preserve every failure

<PARSED TEXT FOR PAGE: 39 / 40>
المراجع العلمية الخارجية الأساسية REFERENCES

Role Source ID

fine-grained attribution and leave-one-out chains Chen et al. 2026 - Long-Horizon Agent Trajectory Attribution -. arXiv:2608.06909 E05

automated red teaming, self-play, held-out attacks, robustness OpenAI 2026 - GPT-Red E12.

multi-agent utility can rise with misalignment Shen et al. 2026 - AI Organizations are More Effective but Less. Aligned - arXiv:2604.10290 E13

identity, authorization, auditing and non-repudiation NIST NCCoE 2026 - Software and AI Agent Identity and. Authorization E14

يُعاد البحث الأدبي قبل claim أسبقية أو submission خارجي. كل result benchmark خارجي result reported لا replication داخلية.

<PARSED TEXT FOR PAGE: 40 / 40>
الحكم البحثي وحدود الورقة VERDICT

الخلاصة: تربط الورقة corpus الوعي الوظيفي والـmodel-self وSpace Governance وCapture Reverse بأبحاث alignment agent-multi وهوية الوكلاء. التركيز ليس على وصف التنظيم، بل على قياس التدفق السببي للسلطة، هشاشة المحاذاة الجماعية، الحد الأدنى من self-model، وآليات الميتاحاكمية زمن appeal/repair.

القيمة العلمية ليست زيادة ontology، بل hypotheses يمكن رفضها. إذا لم تحسن التشخيص أو الإسناد أو recovery أو transfer أو evolution مقارنة بالـbaselines الأبسط، يجب دمجها أو إسقاطها. grounded-external

PROMOTE only after mechanism + trace + causal gain + independent evidence + transfer

Paper status: N0-N2 theoretical closure specification. No empirical AGI/ASI claim.