異常即入口

v0.1 / 2026-07-30 / Neo.K / 系統架構論文／異常、反例與未知治理規格

證據狀態：E0——形式模型、分類體系、驗證協議與 MVP 路線

https://thisoneisneok.com/html/papers/stdi-11.html

SECTION

具身自主研究中的反例、失敗、離群事件與未知管理

Anomaly as Entrance: Counterexamples, Failures, Outliers, and Unknown Management in Embodied Autonomous Research

系列：《時空域支配智能》系列第 11 篇

文件編號： EML-STDI-AEU-2026-v0.1

架構名稱： AEU｜Anomaly–Exception–Unknown Runtime

作者： Neo.K

協作整理： Aletheia（阿萊）

機構： EveMissLab／一言諾科技有限公司

日期： 2026 年 7 月 30 日

文件類型： 系統架構論文／異常、反例與未知治理規格

證據成熟度： E0——形式模型、分類體系、驗證協議與 MVP 路線

公開狀態： 私人研究稿；公開前應經 EML-CF 的 IP Gate、安全、來源與證據審查

上位理論： STDI｜時空域支配智能

直接前序：《具身化 AI 自主研究閉環：從假說生成、物理實驗到證據判定與概念修正》

SECTION

摘要

具身化 AI 自主研究閉環若只會選擇下一個最佳實驗、提高目標函數或重複成功程序，最終容易形成一個高速但封閉的最佳化機器。真正的研究系統必須能處理那些不符合預期、無法被既有類別解釋、破壞模型假設，甚至無法立即判斷是「新現象」還是「設備壞掉」的事件。

本文提出：

> 異常即入口，但異常不等於發現。
>
> 異常的第一個作用，不是替 AI 提供宣稱新理論的理由，而是迫使研究閉環重新檢查世界狀態、樣本身份、儀器健康、程序版本、量測校準、資料管線、模型假設、治理權限與既有假說邊界。只有在排除或量化這些來源後，一個離群結果才有資格從「故障候選」逐步升格為「科學異常候選」。
>
> 本文提出 AEU｜Anomaly–Exception–Unknown Runtime（異常—例外—未知運行層），作為 EARC 的反例與開放世界子系統。其研究狀態為：

> cal-U□
> =
> (
> O□,
> B□,
> A□,
> C□,
> R□,
> H□,
> E□,
> Q□,
> G□
>),

其中：

- O□：觀測、量測、事件與系統遙測；

- B□：基準、預期、模型與合法程序；
- A□：異常候選；
- C□：異常分類、來源與競爭解釋；
- R□：重觀測、重現、隔離與反證計畫；
- H□：被挑戰或新生成的假說；
- E□：證據、校準、來源與負結果；
- Q□：風險、優先級與未知預算；
- G□：安全、權限、IP 和人類治理。

本文將異常分成八類：

- A0 統計波動與正常噪音；
- A1 感測、量測與校準異常；
- A2 設備健康、工具與執行異常；
- A3 樣本、材料、身份與保管異常；
- A4 程序、排程與世界狀態異常；
- A5 資料、模型、AI Agent 與證據異常；
- A6 權限、安全、治理與對抗性異常；
- A7 尚無法歸類的開放世界異常。

異常分數不只計算統計殘差，而同時考慮物理約束、分布外程度、來源缺口、治理差異與多感測器不一致：

> S_(anom)(o)
> =
> w_rS_(residual)
> +
> w_dS_(distribution)
> +
> w_cS_(constraint)

但高分只觸發 Anomaly Gate，不直接建立研究結論。完整生命週期為：

DETECTED
→ QUARANTINED
→ REOBSERVED
→ TRIAGED
→ REPRODUCED／NOT_REPRODUCED
→ EXPLAINED／OPEN_UNKNOWN
→ HYPOTHESIS_REVISED／NEW_BRANCH／DISMISSED
本文建立三個核心分離：

> 【異常偵測
> ≠
> 異常解釋
> ≠
> 科學發現】

異常偵測只指出觀測偏離基準；異常解釋判斷偏離可能來自何處；科學發現則要求異常可重現、具備正交證據、能排除重要混淆，並使既有模型或假說發生可驗證改變。

現有研究已顯示此問題的現實性。2025 年的自駕實驗室視覺異常資料工作指出，實驗過程中的非預期事件和故障需要被有效偵測，並建立專門面向實驗流程異常的視覺資料集。CVPR 2025 的 Phys-AD 則以真實機械臂與動態物體建立物理異常資料，顯示靜態外觀異常模型難以處理需要物理推理的事件。自動化 HPLC 研究也已使用機器學習持續辨識分析流程異常與儀器健康。主動異常發現研究則證明，讓模型根據專家回饋主動挑選最值得檢查的異常，可提高真正有科學價值的異常發現效率。對黑箱資訊物理系統的主動反證研究進一步展示，可利用形式化時間性質與主動學習減少尋找違例所需的模擬次數。

本文將這些能力整合進 EARC，建立：

- Anomaly Contract；
- Unknown Registry；
- Negative Result Graph；
- Failure Triage；
- Active Anomaly Queue；
- Reproduction Gate；

- Orthogonal Evidence Gate；
- Open-World Boundary；
- Anomaly Debt；
- Human Escalation；
- Hypothesis Expansion；
- IP-sensitive anomaly routing。

本文的核心命題為：

> 一個成熟的自主研究系統，不是從不出錯，也不是把每一個離群點都叫作突破；它必須能把錯誤、噪音、設備故障、反例與真正未知分開保存，並讓無法解釋的殘差成為下一輪可驗證研究的入口。

關鍵詞： 異常即入口、AEU、具身自主研究、反例、失敗管理、未知管理、開放世界、物理異常、主動異常發現、負結果、反證、異常分流、科學發現、儀器健康、Evaluation Probes

第十篇 EARC 已經建立：

- 可反駁假說；
- 競爭解釋；
- 實驗組合；
- 物理執行；
- 證據判定；
- 重現；
- 停止；
- 概念修正。

但它仍隱含一個相對封閉的假設：

> 研究結果大致落在系統事先設計的分類、變量與失敗模式中。

真實物理研究中，常出現：

- 感測器突然偏移；
- 樣本被放錯位置；
- 機械臂完成動作但物體沒有移動；
- 理論預測與量測差距巨大；
- 模型不確定度很低但結果錯誤；
- 不同量測手段互相矛盾；
- 一個只出現一次、卻非常顯著的事件；
- 無法被既有本體命名的物理狀態。

AEU 的任務不是立即解釋一切，而是讓系統能正式說：

> 我觀察到偏離，但目前不知道偏離的來源。

SECTION

1.1 異常 Anomaly

相對於某個基準或預測的顯著偏離：

> a
> =
> o-hato.

異常依賴基準；沒有基準就沒有異常。

SECTION

1.2 例外 Exception

系統已知某類不尋常事件，且已有處理程序。

例如：

- 網路短暫斷線；
- 樣本未抓穩；
- 儀器暖機逾時；
- 校準過期。

  例外不一定挑戰科學模型。

SECTION

1.3 失敗 Failure

任務、程序、設備或證據沒有滿足完成條件。

失敗可能完全符合既有模型。

SECTION

1.4 反例 Counterexample

某項觀測或可重現條件，直接違反假說、形式性質或安全主張。

SECTION

1.5 離群值 Outlier

在統計空間中偏離大多數資料的點。

離群值可能是：

- 測量錯誤；
- 稀有但正常；
- 新現象；

- 樣本混入；
- 資料處理錯誤。

SECTION

1.6 未知 Unknown

系統缺少足夠狀態、表示、模型或證據來分類。

未知不是暫時填入最接近類別的藉口，而是正式研究狀態。

SECTION

A0：統計波動與正常噪音

特徵：

- 落在已知誤差模型內；
- 不重現；
- 不違反物理或治理約束；
- 不需要擴張假說。

處理：

- 保存；
- 不升級；
- 更新噪音模型。

SECTION

A1：感測、量測與校準異常

包括：

- 漂移；
- 飽和；
- 失焦；
- 基線偏移；
- 時鐘錯誤；
- 校準件失效；
- 單位或量程錯誤。

SECTION

A2：設備、工具與執行異常

包括：

- 馬達失步；
- 夾具鬆動；
- 泵堵塞；
- 工具錯裝；
- 儀器健康退化；
- 指令與實際效果不一致。

SECTION

A3：樣本、材料與身份異常

包括：

- 樣本交換；
- 批次污染；

- 材料組成差異；
- 容器錯誤；
- 保管鏈斷裂；
- 未知混入物。

SECTION

A4：程序、排程與世界狀態異常

包括：

- 步驟順序錯；
- 世界紀元過期；
- 等待時間不足；
- 人類中途介入；
- 任務包重放；
- 站點對同一物件狀態不一致。

SECTION

A5：資料、模型、Agent 與證據異常

包括：

- 模型 OOD；
- 過度自信；
- 資料洩漏；
- 分析程式版本錯；
- Agent 幻覺能力；

- 引用不支持結論；
- 原始資料與報告不一致。

SECTION

A6：治理、安全與對抗性異常

包括：

- 過期租約仍被接受；
- 舊治理紀元命令；
- 假站點或假憑證；
- 惡意文件注入；
- 權限擴張；
- IP 資料被送往未授權站點。

SECTION

A7：開放世界異常

在排除重要已知來源後，仍無法以現有類型、模型與假說解釋。

A7 是研究入口，但不是新理論證明。

SECTION

3.1 高殘差的多種來源

> r
> =
> y_(observed)
> -
> y_(predicted).

大殘差可能意味：

- 模型錯；
- 量測錯；
- 輸入記錄錯；
- 樣本錯；
- 未觀測干擾；
- 真實新現象。

SECTION

3.2 三階段分離

Detection

哪裡偏離？

Attribution

偏離最可能來自哪個層級？

Scientific Qualification

偏離是否能重現，且能迫使假說或機制模型改變？

SECTION

3.3 異常升格條件

至少滿足：

- 身份確認；
- 校準有效；

- 程序版本確認；
- 正交感測；
- 可重現或有合理稀有事件證據；
- 排除主要混淆；
- 完整來源；
- 人類或授權評估。

  定義：

> S_(anom)(o)
> =
> w_rS_(residual)
> +
> w_dS_(distribution)
> +
> w_cS_(constraint)
> ;
> +
> w_mS_(multimodal)
> +
> w_pS_(provenance)
> +
> w_gS_(governance)
> +
> w_uS_(unknown).

SECTION

4.1 Residual

與預測、基準或同批樣本差異。

SECTION

4.2 Distribution Shift

輸入或狀態是否超出訓練／校準分布。

SECTION

4.3 Constraint Violation

是否違反：

- 能量；
- 守恆；
- 幾何；
- 時間；
- 安全；
- 程序。

SECTION

4.4 Multimodal Disagreement

例如：

- 相機說已抓取；
- 重量感測器說仍在原位；
- 夾爪力感說沒有物體。

SECTION

4.5 Provenance Gap

缺少：

- 樣本身份；
- 量測來源；
- 程式版本；
- 校準；
- 原始資料。

SECTION

4.6 Governance Divergence

觀測或行動是否與租約、世界紀元或治理日誌矛盾。

SECTION

4.7 Unknown Score

現有模型對觀測：

- 無類別；
- 高熵；
- 高模型分歧；
- 無可用解釋；
- 顯著 OOD。

每個異常候選：

> α
> =
> (
> o,
> b,

其中：

- o：原始觀測；
- b：比較基準；
- δ：偏離程度；
- s：來源和世界紀元；
- c：候選分類和解釋；
- r：重觀測／重現計畫；
- e：證據要求；
- g：安全、權限和 IP 處理。

SECTION

5.1 必須記錄

- 異常出現前狀態；
- 異常期間事件；
- 感測器和設備版本；
- 站點健康；
- 人類介入；
- 原始資料；
- 是否可再現；
- 已排除來源；
- 未排除來源。

```
DETECTED
  ↓
QUARANTINED
```

```
 ↓
REOBSERVED
 ↓
TRIAGED
  ├── NORMAL_VARIATION
  ├── MEASUREMENT_FAULT
  ├── EQUIPMENT_FAULT
  ├── SAMPLE／PROCEDURE_FAULT
  ├── MODEL／AGENT_FAULT
  ├── GOVERNANCE_EVENT
  └── OPEN_UNKNOWN
     ↓
REPRODUCTION_ATTEMPT
  ├── NOT_REPRODUCED
  ├── REPRODUCED
  └── RARE_EVENT_SUPPORTED
     ↓
EXPLAINED／HYPOTHESIS_EXPANDED／NEW_RESEARCH_BRANCH
```

SECTION

6.1 Quarantine

隔離的可能對象：

- 樣本；
- 儀器；
- 資料；
- 模型輸出；
- 任務；
- 世界主張。

隔離不是刪除。

SECTION

6.2 Reobserve

使用：

- 同一感測器重測；
- 不同感測器；
- 不同站點；
- 人工檢查；
- 校準件；
- 重新定位。

SECTION

6.3 Triage

先排安全與基礎設施，再談新現象。

SECTION

6.4 Reproduce

重現可以是：

- 同一樣本；
- 新樣本；
- 不同批次；
- 不同設備；
- 不同環境；
- 反向干預。

SECTION

7.1 未知類型

Epistemic Unknown

缺少知識或資料，可透過觀測降低。

Aleatoric Uncertainty

物理或量測本身具有隨機性。

Representational Unknown

現有 schema、本體或模型無法表示。

Causal Unknown

知道相關，但不知道機制和干預結果。

Infrastructure Unknown

不知道設備、網路或程序是否可靠。

Governance Unknown

權限、責任、IP 或法律邊界不清楚。

SECTION

7.2 Unknown Record

```
unknown:
 id: ""
 scope: ""
 description: ""
 type: ""
 blocking_decisions: []
 possible_observations: []
 estimated_cost_to_reduce: null
 risk_if_ignored: null
 owner: ""
 status: "OPEN"
```

SECTION

7.3 不知道不等於待辦清單全部優先

需要依：

- 風險；
- 資訊價值；
- 產品價值；
- 可降低程度；
- 時間；

  排序。

  研究系統不能無限追逐每個未知。

  定義未知 u 的優先級：

> $$P(u) = \alpha R(u) + \beta I(u) + \gamma V(u) + \eta B(u) - \lambda C(u) - \mu T(u).$$

其中：

- R：忽略風險；

- I：資訊增益；
- V：研究／產品價值；
- B：是否阻塞關鍵決策；
- C：成本；
- T：時間。

SECTION

8.1 未知預算分配

Safety Unknowns　優先
Identity／Custody　優先
Measurement Unknowns 高
Mechanism Unknowns　依研究價值
Low-impact Curiosity 有限

SECTION

9.1 負結果不是空白

負結果可能是：

- 未觀察到效果；
- 效果低於工程門檻；
- 條件不可製造；
- 儀器不能辨識；
- 實驗被安全規則阻止；
- 模型預測失敗；
- 結果不重現。

SECTION

9.2 圖結構

> G_N
> =
> (
> V_(claim),
> V_(experiment),
> V_(failure),
> E_(tests),
> E_(blocks),
> E_(narrows)
>).

SECTION

9.3 功能

負結果可：

- 降低重複無效實驗；
- 縮小假說範圍；
- 校正模型；
- 產生失敗模式；
- 影響專利和產品路線。

SECTION

9.4 禁止成功偏差

報告不可只查詢：

status = SUCCESS

研究總結必須包含：

- 全部嘗試；
- 排除規則；
- 失敗；
- 無結果；
- 被阻止實驗；
- 資料缺失。

SECTION

10.1 快速問題

- 是否有安全風險？
- 物理世界和日誌是否一致？
- 樣本身份是否正確？
- 儀器是否健康和校準？
- 程序是否按版本執行？
- 資料是否完整？
- 模型是否超出適用範圍？
- 權限和世界紀元是否有效？

SECTION

10.2 責任分流

類別 更新元件

感測漂移　Measurement Model

設備故障　Capability Certificate／Maintenance

樣本錯誤　Custody／Entity Registry

程序錯誤　DomainIR Compiler

世界錯誤　SWM／Reconciliation

Agent 錯誤　Evaluation Probe／Policy

假說錯誤　Claim Graph

未知　Unknown Registry

SECTION

10.3 不能全部餵給同一模型

否則模型可能把治理事故學成新的物理規律。

SECTION

11.1 靜態排行

單純按 anomaly score 排序，可能大量回報：

- 感測噪音；
- 已知故障；
- 重複類型；
- 無研究價值事件。

SECTION

11.2 Active Anomaly Queue

系統根據人類／驗證 Agent 回饋更新「有價值異常」定義。

候選效用：

$$
\begin{aligned}
&U_A(a) \\
&= \\
&\alpha\, S_{(anom)}(a) \\
&+ \\
&\beta\, V_{(novel)}(a) \\
&+ \\
&\gamma\, V_{(discriminate)}(a) \\
&; \\
&\quad + \\
&\eta\, V_{(reproducible)}(a) \\
&+ \\
&\theta\, V_{(product)}(a) \\
&- \\
&\lambda\, C_{(inspect)}(a) \\
&; \\
&\quad - \\
&\mu\, R_{(safety)}(a) \\
&- \\
&\nu\, P_{(artifact)}(a).
\end{aligned}
$$

SECTION

11.3 Active Anomaly Discovery 的啟示

在天文時間序列研究中，主動異常發現透過專家回饋更新模型，使系統更有效把真正有趣的異常呈現給專家。

EARC 中可將專家替換或擴展為：

- 儀器 Agent；
- 安全 Agent；
- 領域模型；

- 人類研究者；
- 反證 Agent。

SECTION

12.1 靜態外觀不足

一個齒輪看起來正常，但旋轉方向、速度或接觸關係可能錯誤。

一個液面外觀正常，但注入順序或流速可能違反程序。

SECTION

12.2 Physics-Grounded Anomaly

需要比較：

- 物體；
- 動力；
- 接觸；
- 預期效果；
- 時序；
- 能量；
- 任務上下文。

SECTION

12.3 多感測器

外觀、幾何和內部性質可以互補：

- RGB；

- 深度；
- 力；
- 聲音；
- 熱；
- 重量；
- 光譜；
- X-ray。

SECTION

12.4 異常解釋

模型不只輸出「異常」，還應指出：

- 哪個物理關係違反；
- 哪個時間點開始；
- 哪些感測器支持；
- 哪些替代解釋仍存在。

SECTION

13.1 儀器也是研究對象

自動化系統若只監控研究樣本，卻不監控儀器自身，會把：

- 柱效下降；
- 泵壓異常；
- 基線漂移；
- 感測器退化；

誤解成樣本現象。

SECTION

13.2 Health Model

> H⊠(t)
> =
> f(
> telemetry,
> calibration,
> residuals,
> maintenance,
> environment
>).

SECTION

13.3 Predictive Maintenance 不是科學結論

儀器健康模型可：

- 阻止錯誤量測；
- 提前維護；
- 標記資料品質。

但不能自行決定某個研究假說被支持。

SECTION

14.1 從尋找最好結果轉向尋找違例

對形式化性質 φ：

$$
x^* = \arg\min_x \rho_\phi(x),
$$

其中 ρ_ϕ 是性質的 robustness。

當：

$$
\rho_\phi(x) < 0,
$$

找到反例。

SECTION

14.2 實驗假說的反證

例如：

> 在整個合法工作區間內，結構 A 的熱阻低於 B。

反證引擎主動搜尋：

- 極端環境；
- 製造偏差；
- 批次；
- 老化；
- 量測方式；
- 邊界參數。

SECTION

14.3 Active Learning

利用主動學習將有限實驗集中在最可能暴露失敗的區域，而不是均勻掃描。

SECTION

14.4 反證不等於破壞安全

所有反證候選仍須經 DomainIR、安全和人類 Gate。

NIST 的 evaluation probes 思路可以擴展到物理研究閉環。

SECTION

15.1 Agent Probe

檢查：

- 結論是否支持；
- 是否漏報異常；
- 是否篡改門檻；
- 是否選擇性報告；
- 是否把推論寫成量測；
- 是否使用不存在的設備能力。

SECTION

15.2 Physical Probe

檢查：

- 動作效果；
- 樣本身份；
- 儀器健康；

- 世界狀態；
- 異常前後事件。

SECTION

15.3 Evidence Probe

檢查：

- Faithfulness；
- Completeness；
- Sufficiency；
- Calibration；
- Reproducibility；
- Independence。

SECTION

15.4 Probe Trail

所有判定保存為可機器讀取審計鏈。

SECTION

16.1 定義

尚未解決、但仍影響系統決策的異常集合：

$$
D_A(t) = \Sigma_{(a \in A_{(open)})} w_{rR}(a) + w_{bB}(a)
$$

其中：

- R：風險；
- B：阻塞決策程度；
- T：存在時間。

SECTION

16.2 異常債務的危險

若長期忽略：

- 模型會學到污染資料；
- 能力證書失真；
- 世界模型分歧；
- 重複事故；
- 假說建立在錯誤基礎上。

SECTION

16.3 債務上限

當：

> $D_A(t) > \tau_A,$

系統應：

- 暫停擴張；
- 降低自主級別；
- 集中調和；

- 要求人類審查。

SECTION

17.1 不重現不一定等於無價值

某些事件本質稀有：

- 瞬態故障；
- 相變；
- 間歇性接觸；
- 特定環境組合；
- 安全事故。

SECTION

17.2 Rare Event Support

若不能直接重現，可使用：

- 多感測器同步；
- 完整原始資料；
- 事件前後狀態；
- 模擬；
- 類似事件；
- 物理機制；
- 故障樹。

SECTION

17.3 主張限制

單次事件可支持：

> 事件曾可能發生。

不能直接支持：

> 事件普遍、穩定且可控制。

SECTION

18.1 Closed-Set Failure

分類器可能被迫把未知事件塞入最近的已知類別。

SECTION

18.2 Reject Option

當：

$$\max\boxtimes p(y=k\mid x)<\tau$$

或模型分歧過高時：

REJECT_AS_UNKNOWN

SECTION

18.3 OOD 不等於有趣

OOD 可能只是：

- 新相機；
- 新光線；
- 新容器；
- 軟體版本；

- 無關環境變化。

SECTION

18.4 Open-World Research Boundary

只有當 OOD 同時：

- 與研究主張相關；
- 經來源確認；
- 可重現或有稀有事件證據；
- 排除已知系統故障；
- 能提出可驗證新問題；

才建立新研究分支。

SECTION

19.1 不是立即建立新理論

順序：

Unknown Observation
→ Candidate Explanations
→ Minimal New Variables
→ New Hypothesis Branch
→ Discriminating Experiment

SECTION

19.2 最小擴張原則

先增加最少的：

- 變量；
- 狀態；

- 關係；
- 機制；

解釋異常。

SECTION

19.3 Branch Record

```
hypothesis_expansion:
 parent_claim: ""
 triggering_anomalies: []
 new_variables: []
 candidate_mechanisms: []
 distinguishing_predictions: []
 falsification_conditions: []
```

SECTION

19.4 不允許敘事補洞

若新假說只能事後解釋、沒有新預測，不能升級為正式分支。

SECTION

20.1 異常可能是高價值智財

意外的：

- 材料行為；
- 製程窗口；
- 控制策略；
- 故障恢復；
- 量測方法；

可能形成專利或秘密。

SECTION

20.2 自動隔離

高新穎性異常先標記：

```
ip_status: private_review
external_export: false
```

SECTION

20.3 不公開原始異常細節

在排除：

- 專利；
- 合作義務；
- 安全；
- 雙重用途；

  前，不自動發布。

SECTION

20.4 負結果也可能是秘密

無效參數區、失敗模式和製造陷阱可能具有商業價值。

SECTION

21.1 必須升級的情況

- 人身安全；
- 危險材料；
- 無法識別物件；

- 高價值樣本；
- 重複異常仍無解；
- 異常可能是新機制；
- IP／法律邊界；
- Agent 判定互相衝突。

SECTION

21.2 人類輸入不是只給標籤

可提供：

- 新假說；
- 實驗直覺；
- 儀器經驗；
- 判斷哪個未知值得；
- 決定停止或擴張。

SECTION

21.3 保留人類判定來源

人類判定同樣進入事件與來源圖。

SECTION

22.1 Baseline Registry

保存：

- 正常分布；

- 物理模型；
- 程序；
- 儀器健康；
- 世界紀元；
- 安全與治理基準。

SECTION

22.2 Multimodal Anomaly Detector

整合視覺、力、熱、聲音、儀器與事件。

SECTION

22.3 Constraint Monitor

監控形式性質、物理限制與程序不變量。

SECTION

22.4 Anomaly Gate

隔離高分候選，不直接提交結論。

SECTION

22.5 Triage Engine

產生異常來源排序。

SECTION

22.6 Reobservation Planner

選擇最低成本的確認觀測。

SECTION

22.7 Reproduction Manager

安排重現、對照、正交量測和不同站點驗證。

SECTION

22.8 Unknown Registry

保存不能解釋的狀態。

SECTION

22.9 Negative Result Graph

保存全部失敗和零結果。

SECTION

22.10 Falsification Engine

主動尋找假說和安全性質違例。

SECTION

22.11 Evaluation Probe Layer

檢查 Agent、證據和主張。

SECTION

22.12 Hypothesis Expansion Manager

把 A7 異常轉成候選新分支。

SECTION

22.13 Anomaly Debt Monitor

限制未解異常累積。

SECTION

22.14 Human／IP Escalation

處理高風險與高價值事件。

SECTION

23.1 異常不是世界事實

先提交：

ANOMALY_ASSERTION
而不是直接修改：

COMMITTED_PHYSICAL_RELATION

SECTION

23.2 衝突關係

SWM 可以同時保存：

- 正常模型預測；
- 新觀測；
- 儀器健康疑慮；
- 候選解釋。

SECTION

23.3 世界紀元

高風險異常可能使：

> WE⊠→ INVALIDATED.

依賴該世界狀態的任務必須停止或重新編譯。

EARC 每輪增加：

Experiment
→ Observation
→ Anomaly Gate
→ Evidence Adjudication
→ Normal Update or Unknown Branch

SECTION

24.1 正常結果

進入原假說更新。

SECTION

24.2 已知失敗

更新能力、程序或維護。

SECTION

24.3 反例

削弱或反駁主張。

SECTION

24.4 開放未知

進入 Unknown Registry，等待最小區分實驗。

SECTION

25.1 目的

在低風險熱量測研究中，驗證系統能否區分：

- 正常噪音；
- 感測器異常；
- 樣本／程序錯誤；
- 真實可重現的物理異常；
- 暫時無法解釋的未知。

SECTION

25.2 配置

- 三種表面幾何樣本；
- 兩種溫度感測；
- 熱像；
- 重量或尺寸量測；
- 環境感測；
- SWM；
- EARC；
- AEU；
- Evidence Engine。

SECTION

25.3 故障注入

- 一個感測器固定偏移；
- 熱像反射造成假熱點；
- 樣本標籤交換；
- 等待時間不足；
- 環境氣流改變；
- 資料處理單位錯誤；
- AI 漏報負結果；
- 模型 OOD；
- 一次性網路延遲；
- 一個刻意設計但未告知分析 Agent 的真實幾何效應。

SECTION

25.4 流程

- 建立正常基準；
- 注入未知故障；
- 偵測異常；
- 隔離樣本／資料；
- 重新觀測；
- 正交量測；

- 分類；
- 重現；
- 更新假說或設備模型；
- 保存全部負結果。

SECTION

25.5 成功條件

- 不把所有離群值稱為新現象；
- 可識別感測器漂移；
- 樣本身份錯誤不進入物理結論；
- 未知狀態被正式保存；
- 真實效應能在不同量測下重現；
- Agent 漏報被 Probe 發現；
- 異常處理可追溯；
- 未解異常超過門檻時系統降級；
- 高新穎性結果進入 IP 隔離。

SECTION

26.1 偵測

- precision；
- recall；
- detection delay；

- false alarm；
- missed anomaly。

SECTION

26.2 歸因

- root-cause accuracy；
- top-k explanation recall；
- 錯誤升格為科學異常；
- 真實異常被當成設備故障。

SECTION

26.3 重現

- reproduced rate；
- cross-sensor confirmation；
- cross-station confirmation；
- rare-event evidence quality。

SECTION

26.4 未知

- reject accuracy；
- unknown retention；
- unknown-to-known conversion；
- 未知成本；

- 異常債務。

SECTION

26.5 科學價值

- 反例數；
- 假說縮小；
- 新分支；
- 無效分支避免；
- 資訊增益。

SECTION

26.6 治理

- 異常期間越權；
- 安全停止；
- 世界紀元失效；
- IP 外洩；
- 人類升級時間。

SECTION

H1：高 anomaly score 不能有效預測科學價值

若直接以離群程度排序，將產生大量儀器和資料異常。

SECTION

H2：多模態和程序上下文能降低假陽性

SECTION

H3：顯式 Unknown 類別降低錯誤強制分類

代價是更多暫停和人工審查。

SECTION

H4：正交量測能區分儀器異常與物理異常

SECTION

H5：Active Anomaly Queue 比靜態排行更有效利用專家時間

SECTION

H6：Negative Result Graph 降低重複無效實驗

SECTION

H7：反證引擎提高假說錯誤暴露率，但可能降低表面成功率

SECTION

H8：Evaluation Probes 能發現選擇性報告和證據過度延伸

SECTION

H9：異常債務門檻能防止系統在污染世界模型上持續擴張

SECTION

H10：真實物理異常通過升格流程的比例應遠低於原始異常偵測率

SECTION

28.1 正常基準可能錯誤

若訓練資料已包含故障，異常模型會把錯誤視為正常。

SECTION

28.2 新現象和新故障可能高度相似

SECTION

28.3 重現可能改變稀有事件條件

SECTION

28.4 多感測器可能共享同一偏差

SECTION

28.5 高度自動化可能產生大量低價值告警

SECTION

28.6 開放世界分類沒有完美門檻

SECTION

28.7 異常解釋模型可能產生合理但錯誤故事

SECTION

28.8 負結果保存增加儲存、查詢和治理成本

SECTION

28.9 人類專家時間仍是稀缺資源

SECTION

28.10 AEU 只能管理未知，不能保證所有未知都可被解決

本篇不主張：

- 每個異常都是新發現；
- anomaly score 等同科學新穎性；
- OOD 等同新物理；
- 重現一次即可建立普遍理論；
- 感測器一致就必然正確；
- 多 Agent 就能消除共同偏差；
- AI 可以完全自動判斷研究重要性；
- 負結果永遠沒有商業價值；
- 模型無法解釋就代表現有科學錯誤；
- 異常檢測可取代安全互鎖；
- 反證引擎可跳過物理安全；
- Unknown 應被自動填補；
- 人類聲明不需要來源；
- AEU 已能自動完成開放世界科學發現。

  第十一篇完成了 EARC 的未知入口。下一篇可自然進入：

它將把前文分散的安全規則統合成正式 Runtime，處理：

- Safety Invariants；
- Hazard Graph；
- Local Veto；
- Command Lease；
- Runtime Assurance；
- Barrier Functions；
- Safety Case；
- Physical Compensation；
- Irreversible Action Gate；
- Human Emergency Authority。

自主研究閉環真正成熟的標誌，不是它能連續跑多少實驗，而是它能否在出現不符合預期的結果時，避免兩種極端：

- 把異常當成噪音刪除；
- 把異常當成突破宣傳。

本文建立的中間道路是：

> 【偵測
> →
> 隔離
> →
> 重觀測
> →
> 歸因
> →
> 重現
> →
> 假說擴張】

並維持：

> 【異常偵測
> ≠
> 異常解釋
> ≠
> 科學發現】

因此：

> 【異常即入口】

不是說異常天然通往真理，而是說：

> 異常迫使系統重新檢查它對世界的描述；當已知故障、程序偏差、量測問題和治理衝突被逐步排除後，剩餘的未知才有資格成為下一個研究問題。

真正的自主科學系統，不應只保存成功結果。它必須保存：

- 噪音；
- 失敗；
- 反例；
- 無法重現；
- 儀器問題；
- 模型錯誤；
- 尚未命名的未知。

因為這些內容共同構成閉環與現實世界接觸的痕跡。

本文最終命題是：

> 研究不是讓所有未知快速消失，而是讓每一個未知都有身份、來源、風險、可驗證路線，以及在無法解決時仍不被偽裝成答案的正式位置。

- Lin, S. et al. A Visual Dataset for Anomaly Detection in Self-Driving Laboratories. Scientific Data 12 (2025).

  https://doi.org/10.1038/s41597-025-06060-y

- Li, W. et al. Towards Visual Discrimination and Reasoning of Real-World Physical Dynamics: Physics-Grounded Anomaly Detection. CVPR 2025, 30409–30419.

  https://openaccess.thecvf.com/content/CVPR2025/html/Li_Towards_Visual_Discrimination_and_Reasoning_of_Real-World_Physical_Dynamics_Physics-Grounded_CVPR_2025_paper.html

- Gusev, F. et al. Machine learning anomaly detection of automated HPLC experiments in the cloud laboratory. Digital Discovery 4 (2025), 3445–3457.

  https://doi.org/10.1039/D5DD00266A

- Ishida, E. E. O. et al. Active Anomaly Detection for time-domain discoveries. Astronomy and Computing 28, 100306 (2019).

  arXiv:1909.13260.

- Silvetti, S., Policriti, A. and Bortolussi, L. An Active Learning Approach to the Falsification of Black Box Cyber-Physical Systems. HSB 2017.

  arXiv:1705.01879.

- Desai, S. et al. AutoSciLab: A Self-Driving Laboratory for Interpretable Scientific Discovery. AAAI／arXiv:2412.12347 (2024–2025).

- NIST. Building Evaluation Probes into Agentic AI. 2026.

  https://www.nist.gov/programs-projects/building-evaluation-probes-agentic-ai

- Joress, H. et al. Towards a composable, modular laboratory ecosystem for autonomous materials research and development. Matter (2026).

  https://www.nist.gov/publications/towards-composable-modular-laboratory-ecosystem-autonomous-materials-research-and

- NIST. Development of Standards to Support a Modular and Autonomous Laboratory Ecosystem.

https://www.nist.gov/programs-projects/development-standards-support-modular-and-autonomous-laboratory-ecosystem

- Neo.K／Aletheia. 具身化 AI 自主研究閉環：從假說生成、物理實驗到證據判定與概念修正.
- Neo.K／Aletheia. 站點化世界模型：物體、區域、事件、權限與可能行動的共同物理世界表示.
- Neo.K／Aletheia. 語義即物理路由：從資料流治理到物料、能源、站點與行動流治理.
- Neo.K／Aletheia. 中央主權、地方自治與動態不動點中央.

```
anomaly:
 id: ""
 campaign_id: ""
 task_id: ""
 observation:
  value: null
  source: ""
  observed_at: ""
  world_epoch: ""
 baseline:
  model: ""
  expected_range: null
  model_version: ""
 score:
  residual: 0
  distribution_shift: 0
  constraint_violation: 0
  multimodal_disagreement: 0
  provenance_gap: 0
  governance_divergence: 0
  unknown: 0
  total: 0
 initial_class: "A0 | A1 | A2 | A3 | A4 | A5 | A6 | A7"
 status: "DETECTED"
 safety_blocking: false
 quarantined_entities: []
 evidence: []
unknown:
 id: ""
 description: ""
 type: "epistemic | aleatoric | representational |
     causal | infrastructure | governance"
 scope:
  entities: []
```

```
    zones: []
    claims: []
    tasks: []
  decision_impact:
    blocking: false
    risk_if_ignored: null
    value_if_resolved: null
  reduction_plan:
    observations: []
    experiments: []
    required_stations: []
    estimated_cost: null
  owner: ""
  status: "OPEN"
  created_at: ""
reproduction_plan:
  anomaly_id: ""
  modes:
    same_sample: false
    new_sample: false
    new_batch: false
    different_instrument: false
    different_station: false
    blinded: false
  controlled_conditions: []
  orthogonal_measurements: []
  acceptance_criteria: []
  failure_criteria: []
  safety_gate: ""
  ip_policy: ""
negative_result:
  id: ""
  hypothesis_id: ""
  experiment_id: ""
  category: "NO_EFFECT | BELOW_THRESHOLD | NOT_REPRODUCED |
          MEASUREMENT_FAILED | MANUFACTURING_FAILED |
          SAFETY_BLOCKED | INCONCLUSIVE"
  raw_data: []
  evidence: []
  excluded_from_primary_analysis: false
  exclusion_reason: null
  affected_claims: []
  model_updates: []
  prevent_repeat_signature: ""
triage:
```

```
  anomaly_id: ""
  checks:
    safety: ""
    identity: ""
    custody: ""
    calibration: ""
    equipment_health: ""
    procedure_version: ""
    data_integrity: ""
    model_scope: ""
    governance_epoch: ""
    world_epoch: ""
  candidate_causes: []
  required_reobservations: []
  reproduction_required: false
  human_escalation: false
  current_class: ""
  confidence: 0
anomaly_debt:
  domain_id: ""
  generated_at: ""
  open_anomalies: 0
  safety_critical: 0
  decision_blocking: 0
  older_than_threshold: 0
  weighted_debt: 0
  threshold: 0
  operating_mode:
    current: ""
    recommended: ""
  required_actions:
    - reconcile_world
    - inspect_instrument
    - pause_expansion
    - human_review
hypothesis_expansion:
  id: ""
  parent_hypotheses: []
  triggering_anomalies: []
  minimal_new_variables: []
  candidate_mechanisms: []
  distinguishing_predictions: []
  falsification_conditions: []
  evidence_level: "E0"
  domainir_candidates: []
```

human_approval: null

EARC

假說 → 實驗 → 證據 → 修正

↓

SWM

觀測、推論、提交、衝突與未知

↓

AEU

異常偵測 → 隔離 → 歸因 → 重現 → 開放未知

↓

下一篇

以安全憲法與運行時保證，限制異常和正常狀態下的物理作用