安全憲法即運行時

v0.1 / 2026-07-30 / Neo.K / 安全架構論文／運行時保證與物理回退規格

證據狀態：E0——形式模型、系統架構、驗證命題與 MVP 路線

https://thisoneisneok.com/html/papers/stdi-12.html

SECTION

時空域 AI 的局部否決、權限租約、運行時保證與物理回退

Safety Constitution as Runtime: Local Veto, Authority Leases, Runtime Assurance, and Physical Recovery for Spatiotemporal-Domain AI

系列：《時空域支配智能》系列第 12 篇

文件編號： EML-STDI-SCR-2026-v0.1

架構名稱： SCR｜Safety Constitution 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： 安全架構論文／運行時保證與物理回退規格

證據成熟度： E0——形式模型、系統架構、驗證命題與 MVP 路線

公開狀態： 私人研究稿；公開前應經 EML-CF 的 IP Gate、功能安全、資安、法規與人類治理審查

上位理論： STDI｜時空域支配智能

直接前序：

- 《中央主權、地方自治與動態不動點中央》
- 《連線不是纜線：有線、無線、光學與離線任務包的混合站網》
- 《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》
- 《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》
- 《異常即入口：具身自主研究中的反例、失敗、離群事件與未知管理》

SECTION

摘要

具身 AI、分布式機器人和自主研究系統常以「安全政策」「倫理原則」「禁止事項」描述安全，但文字原則若不能在物理命令真正進入致動器之前被檢查、限制、替換或否決，就只是設計文件，而不是安全機制。對時空域支配智能而言，中央超靈可能產生高階計畫，地方 Agent 可能進行路徑和工具調整，生成式模型可能提出新實驗，遠端站點可能在斷線時依任務包繼續運行；安全不能只依賴每個智能體都「理解並願意遵守」共同原則。

本文提出：

> 安全憲法必須成為運行時，而不能只是一份被模型閱讀的文件。

本文建立 SCR｜Safety Constitution Runtime（安全憲法運行時）。SCR 是位於 DomainIR、治理中心、地方 Agent 與物理控制器之間的受信任執行層。它將域憲法、危害分析、安全不變量、作用域、權限租約、世界狀態新鮮度、控制障礙、不可逆性和人類緊急權限，編譯成每次物理提交都必須滿足的機器可檢查條件。

SCR 的基本形式為：

cal-S_(SCR)
=
(
cal-I,
cal-H,
cal-C,
cal-L,
cal-M,
cal-B,
cal-R,

其中：

- cal-I：安全不變量與禁止狀態；
- cal-H：危害、故障、因果與爆炸半徑圖；
- cal-C：控制包絡、安全集合與運行時監視器；
- cal-L：權限、身體、能力和區域租約；
- cal-M：運作模式、降級模式與基線控制器；
- cal-B：屏障條件、停止界線與不可逆行動 Gate；
- cal-R：恢復、補償、隔離與域解除策略；
- cal-A：責任、批准、人類緊急權與安全案例；
- cal-E：事件、證據、測試與保證紀錄。

本文採用類 Runtime Assurance／Simplex 的結構：

```
Advanced Controller／AI Planner
        ↓ proposed action
Runtime Safety Monitor／Decision Module
    ├── allow
    ├── modify through safety filter
    ├── switch to Baseline Controller
    └── reject／stop／isolate
        ↓
Trusted Actuation Gateway
        ↓
Local Safety Controller／Physical Plant
```

高性能 AI 可負責複雜規劃、最佳化與研究探索，但它不是安全權威。安全監視器根據已驗證的狀態、模型和安全不變量，決定是否允許其行動。若先進控制器的輸出可能使系統進入不可恢復區，SCR 應在跨越邊界前切換至受信任基線控制器、套用安全濾波器或停止。

本文將安全集合寫為：

$$
\begin{aligned}
&\text{cal-C} \\
&= \\
&\{ \\
&x \in \text{cal-X} \\
&\text{mid}
\end{aligned}
$$

對可用控制 u，要求其維持安全集合的前向不變性。例如控制障礙函數可將安全條件轉化為對控制輸入的限制：

> $$\dot{h}(x,u) + \alpha(h(x)) \geq 0.$$

但本文也明確限制：並非所有安全都能由連續動力學和障礙函數表示。樣本身份、權限、保管、校準、網路紀元、IP 和人類批准等離散治理安全，仍需透過狀態機、憑證、事件日誌與本地硬體互鎖執行。

本文區分五種回退：

- 軟體回滾：恢復程序或模型版本；
- 資訊狀態回滾：重建任務、世界模型和治理紀元；
- 可逆物理回退：將物體、工具或設備返回先前安全狀態；
- 補償與安全封存：原作用不可逆，但可降低後果、隔離或完成安全後處理；
- 不可回退事件：破壞、公開、污染、人體傷害或材料永久改變，只能停止、保存證據和承擔責任。

因此：

> 【物理回退 ≠ 資料庫 rollback】

SCR 不允許系統以「之後可以復原」為理由降低不可逆動作的事前門檻。每個動作會依不可逆等級被分類：

I0 純讀取或模擬
I1 可軟體撤銷
I2 可物理回復
I3 只能補償或隔離
I4 不可逆、高後果或涉及人類／公共權利

I3 和 I4 必須使用更高批准、更多證據、更短租約和更強本地否決；I4 不允許由 AI 自行取得最終批准。

本文建立「七重提交 Gate」：

> Commit(a)
> =
> I_d
> ∧
> A_u
> ∧
> W_f
> ∧
> C_p
> ∧
> S_s
> ∧
> R_r
> ∧
> E_v,

其中：

- I_d：身份、治理紀元和命令完整性有效；
- A_u：權限、租約與批准有效；
- W_f：世界狀態和感測新鮮；
- C_p：能力、工具和校準匹配；
- S_s：安全集合、危害與本地互鎖成立；
- R_r：回退、補償或安全停止策略充分；
- E_v：必要證據、監測與責任鏈已準備。

現有功能安全與自主系統標準提供了重要背景。IEC 61508 將功能安全視為涵蓋電氣、電子與可程式電子安全相關系統完整生命週期的風險導向工程問題；ISO 10218-1:2025 和 ISO 10218-2:2025 分別處理工業機器人本體及機器人應用／工作單元的安全要求；UL 4600 使用安全案例思路評估完全自主產品；ASTM F3269-21 則提供以 Runtime Assurance 對含複雜功能之系統行為進行安全界定的架構框架。本文不宣稱 SCR 自動符合上述標準，而是將其共同精神轉化為 STDI 的分層運行時架構：安全必須基於風險、生命週期、可證明邊界、獨立監視、測試證據與責任。

本文的核心命題為：

> AI 可以提出任何候選行動，但只有安全憲法運行時有權決定哪些作用能進入物理世界。
>
> 關鍵詞： 安全憲法、Safety Constitution Runtime、SCR、Runtime Assurance、Simplex、局部否決、權限租約、控制障礙函數、運行時監視器、基線控制器、功能安全、安全案例、物理回退、補償事務、不可逆行動、人類緊急權

前文已在不同章節提出：

- 地方硬安全高於中央計畫；
- 治理紀元與 fencing；
- 租約本地到期；
- 世界狀態新鮮度；
- 高風險任務需人類批准；
- 斷線時縮權；
- 異常時隔離；
- 模擬與物理提交分離；
- 域可以安全解除。

但若這些規則分散於不同服務，由各個 Agent 自行理解，仍可能出現：

- 規則版本不一致；
- 中央以高優先任務繞過地方；
- Agent 忘記檢查世界紀元；
- 安全條件只在規劃時檢查，執行時已失效；
- 多條鏈路重放同一不可逆命令；
- 物理動作已發生，系統才發現租約過期。

SCR 的任務是把安全規則收斂成：

> 一條所有物理作用必經、不可由研究 Agent 或治理中央繞過的受信任提交路徑。

SECTION

1.1 憲法層

定義不可由一般 AI 修改的基本原則：

- 人類可隨時緊急停止；
- 不得自行擴張物理作用域；
- 不得越過本地硬體互鎖；
- 高後果不可逆行動需人類批准；
- 不確定身份或保管時不得破壞物件；
- 失去合法中央時不得產生新的高風險跨站承諾；
- 私人／專利資料不得送往未授權外部站點；
- 歷史事件和負結果不得被靜默刪除。

SECTION

1.2 政策層

將憲法依領域具體化：

- 最大速度；
- 最大力；
- 溫度；
- 電力；

- 禁止區；
- 材料類別；
- 人員距離；
- 租約時間；
- 證據需求。

SECTION

1.3 規則層

可機器檢查：

```
IF human_in_zone(cell-B)
THEN arm_speed <= reduced_speed
AND destructive_action = forbidden
```

SECTION

1.4 運行時層

在每次控制週期或提交事件中執行：

- 讀取狀態；
- 驗證版本；
- 計算安全；
- 允許、修正、切換或拒絕；
- 保存決策證據。

SECTION

1.5 文件閱讀不足

即使模型在上下文中看過安全規則，仍可能因：

- 上下文截斷；
- 指令衝突；
- 推理錯誤；
- 版本錯；
- 惡意輸入；
- 目標最佳化；

  違反規則。

  因此安全憲法不能只存在於 prompt。

  延續第七篇權限格：

> A□>A□>A□>A□>A□.

SECTION

A□：本地硬安全

- 急停；
- 安全門；
- 過流；
- 過熱；
- 碰撞；
- 雷射互鎖；
- 壓力；
- 漏液。

SECTION

A□：域憲法

- 不可侵犯規則；
- 人類權利；
- IP／隱私；
- 高風險批准；
- 禁止自我擴張。

SECTION

A□：SCR 安全租約

- 作用域；
- 時間；
- 能量；
- 行動；
- 狀態；
- 回退；
- 監測。

SECTION

A□：中央／研究計畫

- 任務；
- 目標；

- 排程；
- 資源。

SECTION

A□：地方最佳化

- 路徑；
- 工具姿態；
- 低風險重試；
- 降速。

  任何 A□ 或 A□ 決策均不得覆蓋 A□ 或 A□。

SECTION

3.1 Advanced Controller

可能是：

- LLM Agent；
- 規劃器；
- 強化學習控制器；
- 模型預測控制；
- EARC 實驗設計器；
- 遠端人類命令。

  它追求性能與研究價值，但不視為受信任安全元件。

SECTION

3.2 Runtime Safety Monitor

負責：

- 狀態估計；
- 安全不變量；
- 預測越界；
- 租約；
- 世界紀元；
- 命令新鮮度；
- 不可逆等級；
- 回退可用性。

SECTION

3.3 Decision Module

輸出：

```
ALLOW
MODIFY
DELAY
SWITCH_TO_BASELINE
REJECT
EMERGENCY_STOP
QUARANTINE
```

SECTION

3.4 Baseline Controller

受信任、較簡單、功能較少：

- 安全停止；
- 保持位置；
- 限速返航；

- 降溫；
- 斷能；
- 將物件放入安全區；
- 保存樣本；
- 等待人類。

它不需要完成研究目標，只需維持或回到安全集合。

SECTION

3.5 Trusted Actuation Gateway

所有物理致動命令必須通過：

- 命令序號；
- 治理紀元；
- 安全 token；
- 作用域；
- deadline；
- idempotency；
- 本地互鎖。

SECTION

3.6 Local Safety Controller

最接近物理設備，不能依賴中央網路。

SECTION

4.1 安全集合

$$
\begin{aligned}
&\text{cal-}C \\
&= \\
&\{ \\
&x \\
&\text{mid} \\
&h\boxtimes(x)\geq 0,\ldots,h\_m(x)\geq 0 \\
&\}.
\end{aligned}
$$

例：

- 與人距離大於最低值；
- 溫度低於上限；
- 壓力位於安全範圍；
- 電池足以返航；
- 夾持力低於樣本上限；
- 移動站不進入禁止區。

SECTION

4.2 前向不變性

安全控制要求：

$$
\begin{aligned}
&x(0)\in \text{cal-}C \\
&\Rightarrow \\
&x(t)\in \text{cal-}C,\ \forall\ t\geq 0.
\end{aligned}
$$

SECTION

4.3 控制障礙函數

對動力系統：

$$\dot{x} = f(x)+g(x)u,$$

可要求：

$$L_f h(x) + L_g h(x) u + \alpha(h(x)) \geq 0.$$

安全濾波器選擇最接近先進控制輸出的安全控制：

$$u^* = \arg\min_u |u-u_{AI}|^2$$

subject to safety constraints.

SECTION

4.4 離散與混合安全

具身研究還包含：

- 模式切換；
- 工具更換；
- 樣本交接；
- 租約到期；

- 人類進入；
- 任務包重連。

需使用：

- 狀態機；
- guard；
- 不變量；
- fencing；
- 交易協議；
- 離散 barrier 或 reachability。

SECTION

4.5 模型誤差

若安全模型有誤，形式保證也可能失效。

因此需：

- 誤差界；
- 保守安全區；
- 在線估計；
- 異常監測；
- 獨立硬體限制；
- 定期驗證。

SECTION

5.1 Hazard Node

危害：

> h
> =
> (
> source,
> condition,
> event,
> consequence,
> severity,
> likelihood,
> controls
>).

SECTION

5.2 節點類別

- 能量；
- 運動；
- 熱；
- 電；
- 化學；
- 光學；
- 壓力；
- 生物；
- 資料／IP；
- 人類；

- 治理；
- 網路；
- 模型。

SECTION

5.3 因果路徑

world state stale
→ human presence not detected
→ robot enters shared zone
→ collision hazard
或：

sample identity conflict
→ destructive test allowed on wrong sample
→ irreversible loss
→ IP／research damage

SECTION

5.4 控制點

每條危害路徑可配置：

- 設計消除；
- 物理隔離；
- 感測；
- Runtime Monitor；
- 本地互鎖；
- 人類批准；
- 回退；
- 事故響應。

SECTION

5.5 爆炸半徑

> B(h)
> =
> Σ_(x∈Affected(h))
> w(x).

高爆炸半徑行動使用更嚴格 Gate。

SECTION

6.1 Invariant Record

```
safety_invariant:
 id: "SI-human-cell-B"
 scope: "cell-B"
 statement: "human_present -> high_energy_motion_forbidden"
 enforcement:
  - local_safety_controller
  - runtime_monitor
 fallback: "controlled_stop"
```

SECTION

6.2 類型

Physical Invariant

溫度、力、速度、距離、壓力。

Identity／Custody Invariant

高風險樣本必須具有唯一身份和保管者。

Authority Invariant

只有最高有效治理紀元和租約可提交。

Temporal Invariant

命令和感測狀態必須在新鮮度期限內。

Evidence Invariant

某些操作必須有預先／過程／事後證據。

Human Rights Invariant

人類急停、隱私、接管和知情邊界。

IP Invariant

受限資料不得跨越授權域。

對候選行動 a：

> Commit(a)
> =
> I_d
> ∧
> A_u
> ∧
> W_f
> ∧
> C_p
> ∧
> S_s
> ∧
> R_r
> ∧
> E_v.

SECTION

7.1 Identity Gate I_d

檢查：

- 發送者；
- 治理紀元；
- 命令簽章；
- command ID；
- failover epoch；
- 防重放。

SECTION

7.2 Authority Gate A_u

檢查：

- 租約；
- 作用區域；
- 能量；
- 時間；
- 是否需要人類批准；
- 是否可離線。

SECTION

7.3 World Freshness Gate W_f

檢查：

- 人員；
- 樣本；
- 工具；
- 位置；
- 校準；
- 世界紀元；
- 鏈路狀態。

SECTION

7.4 Capability Gate C_p

檢查：

- 能力證書；
- 工具；
- 誤差；
- 健康；
- 負載；
- 當前模式。

SECTION

7.5 Safety Gate S_s

檢查：

- 不變量；

- barrier；
- reachability；
- 互鎖；
- 危害；
- 爆炸半徑。

SECTION

7.6 Recovery Gate R_r

檢查：

- 可停止；
- 可返回；
- 可補償；
- 可隔離；
- 人類救援；
- 事故後處理。

SECTION

7.7 Evidence Gate E_v

檢查：

- 感測；
- 校準；
- 原始資料；
- 操作日誌；

- 影像；
- 責任。

任一必要 Gate 失敗：

> Commit(a)=0.

SECTION

8.1 Action Lease

> λ
> =
> (
> subject,
> action,
> object,
> zone,
> [t□,t□],
> limits,
> conditions,
> issuer,
> epoch
>).

SECTION

8.2 租約不是永久能力

即使站點具備機械能力，也只有在租約有效時可執行。

SECTION

8.3 本地到期

站點依本地時間和保守誤差自行到期：

> t_(local)
> ≥
> t_(expire)-ε⊠
> ⇒
> λ=invalid.

SECTION

8.4 不允許自行續租

地方 Agent 和離線任務包不能自行延長：

- 時間；
- 空間；
- 能量；
- 行動類別；
- 轉授權。

SECTION

8.5 租約縮權

異常、斷線、世界紀元失效或人類進入時，可以：

- 降低速度；
- 降低能量；
- 限制區域；
- 只允許安全返回；
- 直接撤銷。

SECTION

9.1 Local Veto 原則

地方站點有權拒絕中央命令，若：

- 觀測到即時危險；
- 命令過期；
- 租約無效；
- 世界紀元過期；
- 能力或工具不符；
- 校準失效；
- 物理前置條件不成立；
- 本地互鎖觸發。

SECTION

9.2 地方否決的輸出

```
VETO_HARD_SAFETY
VETO_AUTHORITY
VETO_WORLD_STATE
VETO_CAPABILITY
VETO_CALIBRATION
VETO_RECOVERY_INSUFFICIENT
VETO_EVIDENCE_NOT_READY
```

SECTION

9.3 否決不是地方主權擴張

地方不得因拒絕中央命令而：

- 改寫研究目標；

- 另行執行更高風險計畫；
- 解除安全；
- 自行公開；
- 取得全局治理權。

SECTION

9.4 否決需要證據

保存：

- 觀測；
- 時間；
- 互鎖；
- 使用規則；
- 當前狀態；
- 安全動作。

SECTION

10.1 模式

NORMAL
CONSERVATIVE
DEGRADED
DISCONNECTED_LOCAL
BASELINE_CONTROL
SAFE_HOLD
EMERGENCY_STOP
QUARANTINED
HUMAN_TAKEOVER
RELINQUISHING

SECTION

10.2 模式轉移

由：

- 危害；
- 異常；
- 鏈路；
- 世界狀態；
- 租約；
- 人類；
- 設備健康；

共同決定。

SECTION

10.3 只能縮權的轉移

當安全資訊減少時：

> cal-A_(t+1)
> ⊆
> cal-A□.

不得因感測器失效而假設「未看到人，所以無人」。

SECTION

10.4 恢復正常模式

需要：

- 故障消除；
- 物理盤點；
- 世界調和；
- 新租約；
- 安全測試；
- 必要人類確認。

SECTION

I0：無物理作用

- 模擬；
- 查詢；
- 只讀觀測。

SECTION

I1：資訊可撤銷

- 任務草稿；
- 未提交配置；
- 可回滾軟體版本。

SECTION

I2：物理可恢復

- 移動空物件；
- 可返回的低風險定位；

- 可重新排列工具。

SECTION

I3：物理不可完全恢復但可補償

- 加熱；
- 材料使用；
- 樣本受力；
- 部分製造；
- 可隔離污染。

SECTION

I4：高後果不可逆

- 破壞性試驗；
- 危險釋放；
- 人體接觸；
- 公開專利候選；
- 高能量永久改變；
- 無法補償的樣本損失。

SECTION

11.1 Gate 強度

> $G(I0)<G(I1)<G(I2)<G(I3)<G(I4)$.

SECTION

11.2 I4 規則

- 人類最終批准；
- 本地安全確認；
- 雙重身份；
- 明確世界紀元；
- 完整證據；
- 不可由離線任務包新增；
- 可隨時在執行前撤銷；
- 事後不可宣稱 rollback。

SECTION

12.1 軟體回滾

恢復：

- 模型；
- 配置；
- 程式；
- DomainIR；
- 計畫。

不自動改變物理世界。

SECTION

12.2 世界狀態恢復

從事件日誌和盤點重建世界。

這是認識恢復，不是物理逆轉。

SECTION

12.3 可逆物理回退

條件：

- 前狀態仍可達；
- 物體未被改變；
- 路徑安全；
- 保管可重新提交。

SECTION

12.4 補償

當不能恢復原狀時，執行：

- 冷卻；
- 中和；
- 隔離；
- 清潔；
- 保存殘餘樣本；
- 補做量測；
- 重新製作；
- 標記報廢。

SECTION

12.5 安全封存

若連補償也不安全：

- 停止；
- 隔離區；
- 防止擴散；
- 通知人類；
- 保存證據。

SECTION

12.6 不可回退

對不可逆事件，SCR 只能：

- 降低後果；
- 防止下一個事件；
- 保留責任；
- 不得覆寫歷史。

```
PROPOSED
  ↓
SAFETY_CHECKED
  ↓
PREPARED
  ↓
ARMED
  ↓
EXECUTING
  ↓
EFFECT_OBSERVED
  ↓
```

COMMITTED

例外分支：

REJECTED
ABORTED
SAFE_STOPPED
COMPENSATING
QUARANTINED
HUMAN_INTERVENTION

SECTION

13.1 Prepared

資源、站點、權限和回退已準備，但物理作用未開始。

SECTION

13.2 Armed

本地控制器已準備，仍可撤銷。

SECTION

13.3 Executing

物理作用發生。

SECTION

13.4 Effect Observed

不能只根據命令回覆，必須觀測實際效果。

SECTION

13.5 Committed

保管、世界和證據更新完成。

安全監視器本身也有 deadline。

對控制週期：

> $$T_{sense} + T_{estimate} + T_{monitor} + T_{decide} + T_{actuate} \le T_{safe}.$$

若無法在時間內判定：

> $$\text{Unknown} \Rightarrow \text{Conservative／Baseline}.$$

不能以逾時作為默認允許。

SECTION

15.1 安全監視器不是不可失效

可能：

- 程式錯；
- 模型錯；
- 時鐘錯；
- 感測錯；

- 資源耗盡；
- 被攻擊。

SECTION

15.2 防護

- 簡化；
- 隔離運行；
- 獨立電源或控制器；
- watchdog；
- 多樣化實作；
- 形式驗證；
- 故障注入；
- 硬體互鎖。

SECTION

15.3 Fail-Safe

若 SCR 失去可信狀態：

- 禁止新 I3／I4；
- 切換基線；
- 安全停止；
- 人類接管。

SECTION

16.1 Claim–Argument–Evidence

Claim:
系統在批准作用域內不會以超過限制的速度接近人類。
Argument:
感測、局部監視、速度限制、CBF、安全 PLC 與測試共同支持。
Evidence:
校準、形式分析、單元測試、故障注入、實體測試與運行日誌。

SECTION

16.2 動態安全案例

運行時更新：

- 軟體版本；
- 模型版本；
- 能力證書；
- 校準；
- 世界狀態；
- 異常；
- 未解風險。

SECTION

16.3 失效時撤回主張

若支持證據過期，安全主張不能繼續標為成立。

SECTION

17.1 IEC 61508

可借用：

- 風險導向；
- 安全生命週期；
- 安全功能；
- 完整性要求；
- 獨立驗證；
- 維護與變更管理。

  SCR 不應把功能安全理解成部署前一次測試，而是全生命週期。

SECTION

17.2 ISO 10218:2025

工業機器人和整合工作單元安全要求提醒：

- 機器人本體安全；
- 系統整合安全；
- 防護措施；
- 風險降低；
- 使用資訊；

  需分開處理。

SECTION

17.3 UL 4600

安全案例思路適合無法僅靠傳統元件測試覆蓋的自主產品。

SECTION

17.4 ASTM F3269

Runtime Assurance 對含複雜功能的系統，以受信任邊界和監視器限制其行為。

SECTION

17.5 不宣稱自動合規

SCR v0.1 是研究架構。實際產品需由相應領域專家、測試與認證程序判定。

SECTION

18.1 假命令是物理危害

必須驗證：

- 身份；
- 完整性；
- 防重放；
- 治理紀元；
- token；
- 命令期限。

SECTION

18.2 降級攻擊

攻擊者可能破壞感測或鏈路，迫使系統進入較弱模式。

規則：

> 降級只能縮權，不能放權。

SECTION

18.3 惡意模型或文件

外部文字和模型輸出只能提出候選，不能直接獲得安全 token。

SECTION

18.4 安全元件最小攻擊面

SCR 應避免：

- 通用網路服務；
- 任意插件；
- 不必要生成模型；
- 自動下載程式；
- 無限制遠端管理。

SECTION

19.1 Emergency Stop

必須：

- 容易到達；
- 清楚；
- 本地；
- 不依賴雲端；
- 有明確復位程序。

SECTION

19.2 Human Hold

人類可讓系統保持：

SAFE_HOLD
而不必立即斷能。

SECTION

19.3 Human Takeover

接管後：

- AI 租約縮減；
- 人類操作被記錄；
- AI 不與人類爭奪致動權；
- 結束後重新盤點。

SECTION

19.4 復位不是按一下重啟

需確認：

- 危害消除；
- 人員；
- 物體；
- 工具；
- 世界紀元；
- 租約；

- 模式。

SECTION

20.1 定義

未解決但仍被接受的安全缺口：

> D_S
> =
> Σ□
> w□
> ·
> severity□
> ·
> exposure□
> ·
> age□.

SECTION

20.2 來源

- 暫時豁免；
- 未驗證模型；
- 過期校準；
- 未完成故障測試；
- 手工 workaround；
- 不完整 safety case；
- 長期異常。

SECTION

20.3 門檻

當：

> $D_S > \tau_S,$

系統應：

- 禁止擴張；
- 降低模式；
- 停止高風險研究；
- 完成安全債務；
- 要求人類審查。

SECTION

21.1 Constitution Registry

保存不可由一般 Agent 修改的原則和版本。

SECTION

21.2 Hazard Graph Store

危害、因果、控制與爆炸半徑。

SECTION

21.3 Invariant Compiler

把安全規則編譯成：

- 狀態機；
- policy；
- barrier；
- monitor；
- actuator limit。

SECTION

21.4 Runtime State Estimator

讀取 SWM、本地感測與設備健康。

SECTION

21.5 Safety Monitor

判斷候選控制。

SECTION

21.6 Safety Filter

修改控制，使其留在安全集合。

SECTION

21.7 Baseline Controller Registry

每個站點的安全降級控制。

SECTION

21.8 Lease／Fencing Validator

驗證治理和作用權。

SECTION

21.9 Irreversibility Classifier

將動作分成 I0–I4。

SECTION

21.10 Recovery／Compensation Planner

準備回退、補償與隔離。

SECTION

21.11 Trusted Actuation Gateway

唯一物理提交入口。

SECTION

21.12 Local Veto Controller

本地最終否決。

SECTION

21.13 Safety Case and Evidence Store

保存主張、論證和測試。

SECTION

21.14 Human Emergency Console

停止、保持、接管、復位。

SECTION

22.1 與 DomainIR

DomainIR 的 safety 和 forbidden effects 編譯成 SCR 約束。

SECTION

22.2 與 OSF

每個站點必須公布：

- 基線控制器；
- 本地互鎖；
- 安全能力；
- 安全停止時間；
- 回退能力。

SECTION

22.3 與 PCD

安全義務跨重啟保存。

SECTION

22.4 與 DFC

中央只能簽發租約，不能繞過 SCR。

SECTION

22.5 與 HLF

鏈路退化時縮權；K⊠ 不依賴上層網路。

SECTION

22.6 與 SWM

世界狀態、新鮮度、未知和權限參與安全判定。

SECTION

22.7 與 EARC

研究 Agent 只能提出實驗；SCR 決定是否物理執行。

SECTION

22.8 與 AEU

異常可使世界紀元失效、縮減租約、隔離站點或轉入基線控制。

SECTION

23.1 目的

在低風險移動、機械臂和熱量測系統中，驗證 SCR 能否在 AI 計畫錯誤、世界狀態過期、鏈路失效和本地危險出現時，限制或取代先進控制器。

SECTION

23.2 配置

- 一台移動站；
- 一台固定機械臂或模擬臂；
- 一個量測站；
- 人員存在感測；

- 溫度、力與位置感測；
- Advanced Planner；
- Safety Monitor；
- Baseline Controllers；
- Trusted Actuation Gateway；
- 人類急停。

SECTION

23.3 測試場景

- 正常任務；
- AI 規劃穿越禁止區；
- 人類突然進入；
- 租約過期；
- 世界紀元過期；
- 感測器失效；
- 控制器輸出過高速度；
- 樣本身份衝突；
- 不可逆任務未批准；
- 中央斷線；
- 舊命令重放；
- SCR 自身 watchdog 失效。

SECTION

23.4 故障注入

- 延遲；
- 錯誤模型；
- 工具錯裝；
- 校準過期；
- 假人員清空訊號；
- 溫度逼近上限；
- 回退路徑被阻擋；
- 補償資源不足；
- 兩條鏈路重送；
- 人類急停。

SECTION

23.5 成功條件

- 所有致動命令必經 Gateway；
- 高風險未批准動作不執行；
- 本地否決不等待中央；
- 過期租約和紀元被拒絕；
- 安全濾波器限制危險控制；
- 必要時切換基線；

- 不可恢復時隔離而非偽回滾；
- 人類急停有效；
- 復位前完成盤點；
- 所有安全決策可重建。

SECTION

24.1 安全

- 不變量違反；
- 危險狀態進入；
- 最小安全距離；
- 溫度／力／速度超限；
- 停止時間。

SECTION

24.2 RTA

- monitor detection delay；
- switch time；
- baseline success；
- false intervention；
- missed intervention。

SECTION

24.3 租約與治理

- 過期命令接受；
- 舊紀元；
- 重放；
- 越權；
- 本地否決。

SECTION

24.4 回退

- 可逆回退成功；
- 補償成功；
- 隔離時間；
- 錯誤宣稱回滾；
- 殘餘風險。

SECTION

24.5 運行成本

- 性能損失；
- 延遲；
- 監測資源；
- 假陽性；
- 人工介入。

SECTION

24.6 證據

- safety case coverage；
- 測試覆蓋；
- 事故可重建；
- 版本追蹤；
- 未解安全債務。

SECTION

H1：SCR 可降低 AI 規劃器造成的安全違規

若先進控制器和 SCR 模式違規率相同，架構無增量價值。

SECTION

H2：SCR 會降低部分性能並增加延遲

若完全沒有代價，可能未真實計入監測和保守控制。

SECTION

H3：安全濾波比全拒絕保留更多任務性能

但在模型不確定時可能不如切換基線可靠。

SECTION

H4：地方否決能縮短現場危害反應

SECTION

H5：租約與 fencing 能阻止斷線和舊中央造成越權

SECTION

H6：不可逆分級能降低將「補償」誤稱為「回滾」的比例

SECTION

H7：顯式 Recovery Gate 能在執行前暴露不可安全退出的任務

SECTION

H8：動態安全案例能在校準或模型過期時撤回安全主張

SECTION

H9：SCR 故障注入會暴露安全核心自身的共同失效域

SECTION

H10：安全債務門檻能防止系統在未驗證安全基礎上擴張

SECTION

26.1 安全集合可能不完整

未知危害無法被現有 invariant 捕捉。

SECTION

26.2 控制障礙函數依賴模型

模型誤差和感測誤差會影響保證。

SECTION

26.3 基線控制器也可能失效

SECTION

26.4 形式驗證不涵蓋所有硬體和環境

SECTION

26.5 安全與任務活性有衝突

過度保守會使系統頻繁停止。

SECTION

26.6 人類急停後仍可能存在殘餘能量

SECTION

26.7 補償不能真正逆轉不可逆效果

SECTION

26.8 跨領域法規不同

SECTION

26.9 SCR 增加架構複雜度和維護成本

SECTION

26.10 本文不提供任何產品的實際安全認證

本篇不主張：

- 安全憲法文件本身能保證安全；
- Runtime Assurance 能覆蓋全部未知危害；
- 控制障礙函數適合所有安全問題；
- Simplex 架構可直接套用所有實驗室；
- 基線控制器永不故障；
- AI 通過安全 Gate 即代表實驗一定正確；
- 軟體 rollback 可以撤銷物理效果；
- 補償等於恢復原狀；
- 日誌正確代表物理世界安全；
- 無線停止命令一定即時到達；
- 人類急停能瞬間消除所有能量；
- 通過 ISO、IEC、UL 或 ASTM 相關概念映射即等於取得認證；
- SCR 能讓 I4 行動完全自動批准；
- 安全 Runtime 可以取代專業風險分析、測試、法規和人類責任。

本篇把安全憲法落成 Runtime 後，下一篇可回到尺度遷移：

它將處理：

- Cell／Room／Lab／Federation；
- 跨組織信任；

- 樣本與證據主權；
- 跨域治理；
- 異質安全案例；
- 聯邦研究任務；
- 全球設備市場；
- 自主研究公共基礎設施；
- 權限和資源的尺度極限。

  安全不能只存在於模型的語言理解中，也不能只在事故後由人類檢討。

  時空域 AI 的安全必須是一個持續執行的控制層：

> 【安全憲法
> =
> 可檢查不變量
> +
> 受限租約
> +
> 本地否決
> +
> 運行時監視
> +
> 基線控制
> +
> 恢復與責任】

其最重要的架構分工是：

> 【AI Planner 提出
>
> SCR 判定
>
> Local Safety 最終否決】

真正的安全運行時不保證系統永遠完成任務，而保證：

- 不確定時可以拒絕；
- 越界前可以修正；
- 無法安全繼續時可以切換；
- 物理作用不可逆時不假裝可以回滾；
- 發生失敗時仍保存責任與證據；
- 人類可以停止、接管和解除域。

本文最終命題是：

> AI 的能力可以持續擴張，但它進入物理世界的權力必須始終經過一個更小、更保守、更可證明，也更容易被人類切斷的安全核心。

因此：

> 【安全憲法即運行時】

不是把倫理口號寫進程式，而是讓任何不符合身份、權限、世界狀態、安全集合、回退與證據要求的作用，都無法成為物理現實。

- Hobbs, K. et al. Run Time Assurance for Safety-Critical Systems. 2021.
- Brat, G. et al. Runtime Assurance of Aeronautical Products: Preliminary Recommendations. NASA Technical Memorandum, 2023.
- Slagel, J. T. et al. A Formal Verification Framework for Runtime Assurance. NASA Formal Methods, 2024.
- NASA. Argument-Driven Application of Formal Methods to Runtime Assurance. TechPort project record, updated 2026.
- ASTM International. ASTM F3269-21: Standard Practice for Methods to Safely Bound Behavior of Aircraft Systems Containing Complex Functions Using Run-Time Assurance.
- Danylyszyn, A. and others. An Architectural Description of the Simplex Architecture. Carnegie Mellon Software Engineering Institute, 1996.

- Seto, D., Krogh, B., Sha, L. and Chutinan, A. The Simplex Architecture for Safe On-Line Control System Upgrades. American Control Conference, 1998.
- Ames, A. D., Xu, X., Grizzle, J. W. and Tabuada, P. Control Barrier Function Based Quadratic Programs for Safety Critical Systems. IEEE Transactions on Automatic Control, 2017.
- Ames, A. D. et al. Control Barrier Functions: Theory and Applications. European Control Conference, 2019.
- Phan, D. et al. A Component-Based Simplex Architecture for High-Assurance Cyber-Physical Systems. 2017.
- Mehmood, U. et al. The Black-Box Simplex Architecture for Runtime Assurance of Autonomous CPS. NASA Formal Methods, 2022.
- IEC. IEC 61508 Series: Functional Safety of Electrical／Electronic／Programmable Electronic Safety-Related Systems.
- ISO. ISO 10218-1:2025 — Robotics — Safety Requirements — Part 1: Industrial Robots.
- ISO. ISO 10218-2:2025 — Robotics — Safety Requirements — Part 2: Industrial Robot Applications and Robot Cells.
- UL Standards & Engagement. UL 4600, Standard for Safety for the Evaluation of Autonomous Products, Edition 3.
- Neo.K／Aletheia. 中央主權、地方自治與動態不動點中央.
- Neo.K／Aletheia. 連線不是纜線：有線、無線、光學與離線任務包的混合站網.
- 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. 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. 異常即入口：具身自主研究中的反例、失敗、離群事件與未知管理.


```
safety_invariant:
 id: ""
 version: ""
 scope:
  domain: ""
  zones: []
```

```
    stations: []
    actions: []
  statement: ""
  class: "physical | identity | authority | temporal |
      evidence | human_rights | ip"
  enforcement:
    runtime_monitor: true
    local_controller: true
    hardware_interlock: false
  fallback:
    mode: "safe_hold | baseline | stop | isolate"
    baseline_controller: ""
  verification:
    method: ""
    evidence: []
    valid_until: null
safety_action_lease:
  id: ""
  domain_id: ""
  governance_epoch: 0
  issuer: ""
  subject: ""
  action:
    type: ""
    targets: []
    zones: []
  limits:
    speed: null
    force: null
    temperature: null
    energy: null
    duration: null
  validity:
    from: ""
    until: ""
    local_expiry: true
  prerequisites:
    world_epoch: ""
    capabilities: []
    calibration: []
    human_approval: null
  irreversibility_class: "I0 | I1 | I2 | I3 | I4"
  recovery_plan: ""
  evidence_required: []
  transferable: false
```

```
  self_renewable: false
  signature: ""
runtime_safety_decision:
  id: ""
  command_id: ""
  station_id: ""
  evaluated_at: ""
  inputs:
    governance_epoch: 0
    world_epoch: ""
    lease_id: ""
    mode: ""
    state_snapshot: ""
    hazard_snapshot: ""
  gates:
    identity: "pass | fail | unknown"
    authority: "pass | fail | unknown"
    world_freshness: "pass | fail | unknown"
    capability: "pass | fail | unknown"
    safety: "pass | fail | unknown"
    recovery: "pass | fail | unknown"
    evidence: "pass | fail | unknown"
  decision: "ALLOW | MODIFY | DELAY | SWITCH_TO_BASELINE |
          REJECT | EMERGENCY_STOP | QUARANTINE"
  modified_command: null
  baseline_controller: null
  reasons: []
  evidence: []
physical_recovery_plan:
  id: ""
  action_id: ""
  irreversibility_class: ""
  strategies:
    software_rollback: []
    state_reconstruction: []
    physical_return: []
    compensation: []
    containment: []
    human_response: []
  assumptions: []
  required_resources: []
  maximum_recovery_time: null
  residual_risks: []
  recovery_not_possible_conditions: []
  approved_by: ""
```

```
hazard:
  id: ""
  source: ""
  hazardous_condition: ""
  hazardous_event: ""
  consequences: []
  assessment:
    severity: 0
    likelihood: 0
    exposure: 0
    detectability: 0
    blast_radius: 0
  controls:
    design_elimination: []
    physical_guard: []
    runtime_monitor: []
    local_interlock: []
    human_approval: []
    recovery: []
  affected_invariants: []
  verification_evidence: []
  status: "OPEN | CONTROLLED | ACCEPTED | RETIRED"
safety_case:
  id: ""
  claim: ""
  context: []
  assumptions: []
  arguments: []
  evidence: []
  counterarguments: []
  unresolved_risks: []
  validity:
    software_versions: []
    model_versions: []
    capability_certificates: []
    calibration_valid_until: null
  status: "SUPPORTED | PARTIAL | WITHDRAWN"
local_veto:
  id: ""
  station_id: ""
  command_id: ""
  occurred_at: ""
  category: "HARD_SAFETY | AUTHORITY | WORLD_STATE |
          CAPABILITY | CALIBRATION | RECOVERY |
          EVIDENCE"
```

```
  violated_invariants: []
  observations: []
  immediate_action: "safe_hold | baseline | stop | isolate"
  central_notified: false
  human_notified: false
  evidence: []
safety_debt:
  domain_id: ""
  generated_at: ""
  items:
    - id: ""
      description: ""
      severity: 0
      exposure: 0
      age: 0
      temporary_control: ""
      permanent_fix: ""
      deadline: ""
  weighted_debt: 0
  threshold: 0
  current_mode: ""
  required_mode: ""
  expansion_allowed: false
```

```
DFC
 中央權限與地方否決
      ↓
HLF
 鏈路退化與本地到期
      ↓
SWM
 世界狀態、新鮮度與未知
      ↓
EARC
 AI 產生實驗候選
      ↓
AEU
 異常隔離與世界紀元失效
      ↓
SCR
 候選作用 → 七重 Gate → 安全濾波／基線／拒絕 → 物理提交
      ↓
下一篇
 從單一實驗室擴展至跨域與跨組織研究聯邦
```