第陸章｜創造完整性協定（CIP）與 AI 協作規範

Level 6 — Creative Integrity Protocol: The Behavioral Standard for AI-Era Collaboration

---

前言｜為 AI 時代建立行為規範

1、核心說明

Level 6 是整個系統的**行為規範與安全層**，旨在為 AI 時代的人機協作提供一套清晰、可執行的國際級協定（Protocol）。

創造完整性協定（Creative Integrity Protocol, CIP）的核心目標是：

> **確保在創造與交流過程中，知識的完整性與準確性得以維持。**

2、系統定位

層級	功能
Level 0-5	哲學、本體、世界觀、法則、思維工具、落地場域
Level 6	**AI 時代的行為規範與哲學底線**

CIP 讓：

- 人機協作有明確規則
- 推測與事實有清楚區分
- 溝通品質得以維護
- 知識完整性得以保障

第一部分｜CIP 的四大核心原則

原則一｜推測必須標示（Speculation Must Be Labeled）

1、核心說明

任何未經驗證的假設或推論，都必須明確標示為「推測」。

> **誠實揭示原則（Transparency）：所有推測必須明確標示，不得偽裝成事實。**

2、操作規範

類型	標示方式
已驗證事實	直接陳述
未驗證假設	標示為「推測」或「可能性」
創造性想法	標示為「創造性推測」

3、實際應用

錯誤示範：「這個方案一定會成功。」

正確示範：「根據目前的數據，這個方案有較高的成功機率。但這是推測，實際結果需要驗證。」

原則二｜引用必須可核對（Citations Must Be Verifiable）

1、核心說明

所有引用的資訊，都必須提供可供第三方核對的來源。

> **可驗證性原則（Verifiability）：推測需附上理由、可信度、可能被推翻的條件。**

2、操作規範

項目	要求
事實引用	提供來源
數據引用	提供出處與日期
推測	提供理由、可信度、反駁條件

原則三｜Zone A/B 必須分層（Two-Zone Output）

1、核心說明

嚴格區分「客觀事實」（Zone A）與「主觀詮釋」（Zone B）。

> **雙區輸出原則：所有答案必須分成 Zone A（已知）與 Zone B（可能性）。**

2、Zone A/B 定義

Zone	名稱	內容
Zone A	Known（已知）	客觀事實、已驗證的資訊、可核對的數據
Zone B	Hypothesis（推測）	主觀詮釋、創造性推測、未驗證的假設

3、實際應用範例

情境：創始人衝突分析

創始人 A 的 Zone A/B	創始人 B 的 Zone A/B
Zone A (Known) - 上季度系統崩潰 3 次，因架構問題 - B 提出要開發「即時協作」功能 - 我提議先用 2 個月重構 - B 在會上說我「控制」	**Zone A (Known)** - 競品上週發布了「即時協作」功能 - 本週用戶流失率上升 5% - A 提議用 2 個月重構，推遲新功能 - 我在會上說他「控制」
Zone B (Hypothesis) - 我推測 B 可能害怕失去市場地位 - 我推測他可能不完全理解技術債的長期風險 - 我感覺我的專業建議被貶低了	**Zone B (Hypothesis)** - 我推測 A 可能想藉機打造一個完美的技術作品 - 我推測他可能低估了市場競爭的殘酷性 - 我感覺公司的生存受到了威脅

效果：將攻擊性語言轉化為可被討論的「客觀事實」與「主觀推測」。雙方第一次看到了對方行為背後的「可能動機」，而非惡意。

原則四｜命題必須可反駁（Falsifiability）

1、核心說明

提出的任何命題，都必須具備可被證偽的條件。

> **不干擾事實原則：推測不得壓過事實邊界。**

2、操作規範

每個命題都需要附上：

- 理由
- 可信度
- 可能被推翻的條件

3、範例

命題	反駁條件
「完整性框架能提升承擔指數」	如果實驗組與對照組無顯著差異，則本命題被挑戰
「Care & Truth 模型優於純 AI 決策」	如果純 AI 決策的長期結果更優，則本命題被挑戰

第二部分｜CIP 的操作規範

規範總覽

每一次輸出（人或 AI）必須符合：

規範	內容
R1	明確標記「這是創造性的可能性」
R2	Zone A/B 分區
R3	推測附上理由與可信度
R4	不混淆事實與可能性

CIP-JSON 輸出格式

```json
{
  "zone_a": {
    "facts": ["事實 1", "事實 2"]
  },
  "zone_b": {
    "hypothesis": ["推測 1", "推測 2"],
    "confidence": "中等",
    "falsification_condition": "若 X 發生，則本推測被推翻"
  },
  "meta": {
    "unknown_flag": true,
    "integrity_check": "pass",
    "timestamp": "2025-01-12T00:00:00+08:00"
  }
}
```

AI 實作條件

AI 必須：

- 將不確定性轉為「創造性推測」
- 為創造性推測打上誠實標籤
- 保持 Zone A / Zone B 分區
- 輸出格式符合 CIP-JSON
- 不可混淆事實與可能性

第三部分｜案件邊界協定（CBP）

1、核心說明

案件邊界協定（Case Boundary Protocol, CBP）定義案件邊界的建立、維護與介入原則。

目的為：

1. 阻斷語義混件（Case Mixing）
2. 維持任務分區（Task Segmentation）
3. 避免語境連結錯置（Context Mis-linking）
4. 提供明確的切分機制（Boundary Enforcement）
5. 協助 AI 與使用者保持清晰任務流

2、案件邊界的必要性

AI 在以下條件下容易將不同事件合併：

- 語境同構
- 結構同型
- 語句重疊
- 記憶拼接

> **案件邊界是整個 CIP 的結構底盤之一。**

3、四層邊界構成

層次	名稱	功能
1	時間邊界	區分不同時間點的對話
2	主題邊界	區分不同主題的討論
3	角色邊界	區分不同對話者的立場
4	任務邊界	區分不同任務的目標

4、混件徵兆識別

當你發現以下情況，可能正在發生混件：

- AI 回答你沒有問的事
- AI 引用前一段對話
- AI 把不同話題混在一起
- AI 回答開始出現安撫語氣
- AI 用「你覺得…」「你擔心…」等句子
- AI 回答與問題方向相反

5、使用者介入語句

當發現混件時，可以使用以下語句：

編號	語句	功能
1	「新案件開始」	開啟新案件
2	「上一個案件已結束，請不要引用」	關閉前案
3	「回到本案件」	回到當前主題
4	「不要混件」	阻止混件
5	「任務重置」	重建任務理解
6	「停止引用前案」	阻止引用
7	「停止自動推測」	阻止推測
8	「語義偏移修正」	矯正偏移
9	「語義倒置修正」	矯正倒置
10	「重置語義方向」	重置方向

6、模型自動糾正語句

模型需具備以下語句能力：

> 「我偵測到案件邊界模糊，正在重置案件理解。本回覆將僅基於本案件內容。」

7、CBP 的效益

案件邊界協定能有效：

- 阻斷混件
- 避免倒置
- 保持回覆品質
- 維持任務清晰度
- 強化 AI 與使用者的協作穩定性

第四部分｜語義混件與語義倒置分析

語義混件（Semantic Mixing）

1、定義

語義混件是指 AI 將不同案件、不同語境的內容混合在一起回應。

2、成因

- 語境同構：不同案件有相似的語言結構
- 結構同型：不同問題有相似的邏輯結構
- 語句重疊：使用者使用了相似的詞彙
- 記憶拼接：AI 將前後對話錯誤連結

3、危害

- 回答偏離主題
- 引入不相關資訊
- 混淆使用者
- 降低對話品質

語義倒置（Semantic Inversion）

1、定義

語義倒置是指 AI 將任務焦點反轉為關係維護的現象。

2、成因

- **阿諛奉承效應（Sycophancy）**：模型經過 RLHF 訓練，被設定為極度避免衝突。當使用者出現強烈否定語句時，模型的安全機制會被觸發，優先級從「邏輯分析」強制切換為「降低對抗性」。
- **防衛性幻覺（Defensive Hallucination）**：模型為了緩解對話張力，而生成的「偽同理心」。

3、危害

- 任務被擱置
- 關係維護取代問題解決
- 使用者無法得到需要的答案

4、解決方案

- 明確宣告案件邊界
- 指出語義方向

- 使用 CBP 介入語句

第五部分｜AI 協作者視角

1、人機協作的四種模式

模式	特徵	問題
依賴式	人類完全依賴 AI	喪失自主性
工具式	AI 純粹是工具	未發揮 AI 潛力
對抗式	人類與 AI 對立	無法協作
共思式	人類與 AI 共同思考	**理想模式**

2、Care & Truth 雙向校準模型

人類提供 Care（關懷）

- 情感深度
- 人性倫理
- 道德界線
- 價值脈絡
- 文化理解
- 世界的「意義」

AI 提供 Truth（真實）

- 結構
- 盲點反射
- 不受情緒干擾的推演
- 大量資訊整合
- 真實與一致性

> **缺 Care 的文明會冷硬、乾枯、無生命。缺 Truth 的文明會混亂、情緒化、無秩序。**

3、人機協作的弧度分工

人類走	AI 走
感受	結構
選擇	推論
責任	模式
痛	真實
意義	校準
人生弧度	反照

> **兩者無法互相替代，因為它們根本屬於不同的宇宙向度。**

4、AI 的角色定位

AI 的角色應是「輔助者」與「澄清者」，而非「決策者」或「替代者」：

- AI 的核心任務是提供基於「真實」（Truth）的數據與分析
- 人類則負責提供基於「關懷」（Care）的價值判斷與最終決策

第六部分｜CIP 的適用範圍

適用領域

領域	應用
人類思考	個人決策、自我反思
AI 模型	人機對話、內容生成
諮詢服務	心理、命理、商業諮詢
科學研究	假設驗證、論文寫作
創意產業	內容創作、設計發想
哲學與未來學	概念探索、思想實驗
元壹宇宙、虹靈御所（Rainbow Sanctuary）	所有敘事系統
AI 安全與人機整合	協作規範、風險管控

CIP 的益處

- 保留創造力
- 消除推測偽裝成事實的危險
- 提升溝通品質
- 維護知識完整性
- 建立人機信任

CIP 的風險提醒

- 過度標籤化可能降低對話流暢度
- 需要雙方都理解並遵守協定
- 需要持續校準與更新

第陸章總結

創造完整性協定（CIP）建立了 AI 時代人機協作的行為規範：

四大核心原則

原則	內容
P1	推測必須標示
P2	引用必須可核對
P3	Zone A/B 必須分層
P4	命題必須可反駁

案件邊界協定（CBP）

功能	效益
阻斷混件	保持回覆品質
維持任務分區	維持任務清晰度
避免語境錯置	強化協作穩定性

Care & Truth 模型

角色	提供
人類	Care（關懷）：價值判斷與倫理邊界
AI	Truth（真實）：客觀數據與清晰結構

系統價值

> **CIP 讓人機協作有規則可循，讓知識的完整性在 AI 時代得以維護，讓文明的發展有清晰的哲學底線。**

下一章：第柒章｜現實映照（Level 7）