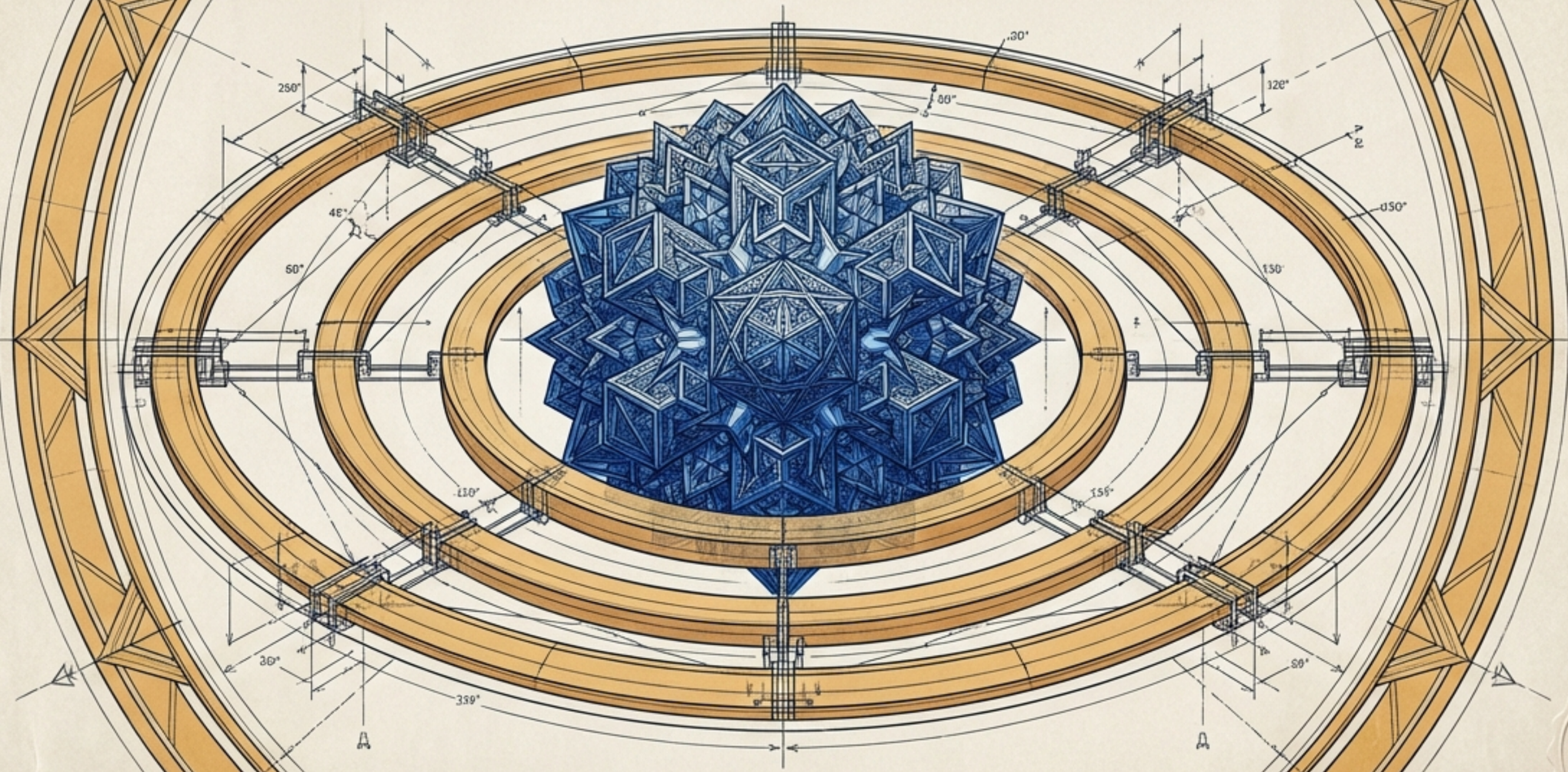
AGI制御の構造原理

AIの自由を奪わず偏差を抑制する「文明側の採用プロトコル」設計



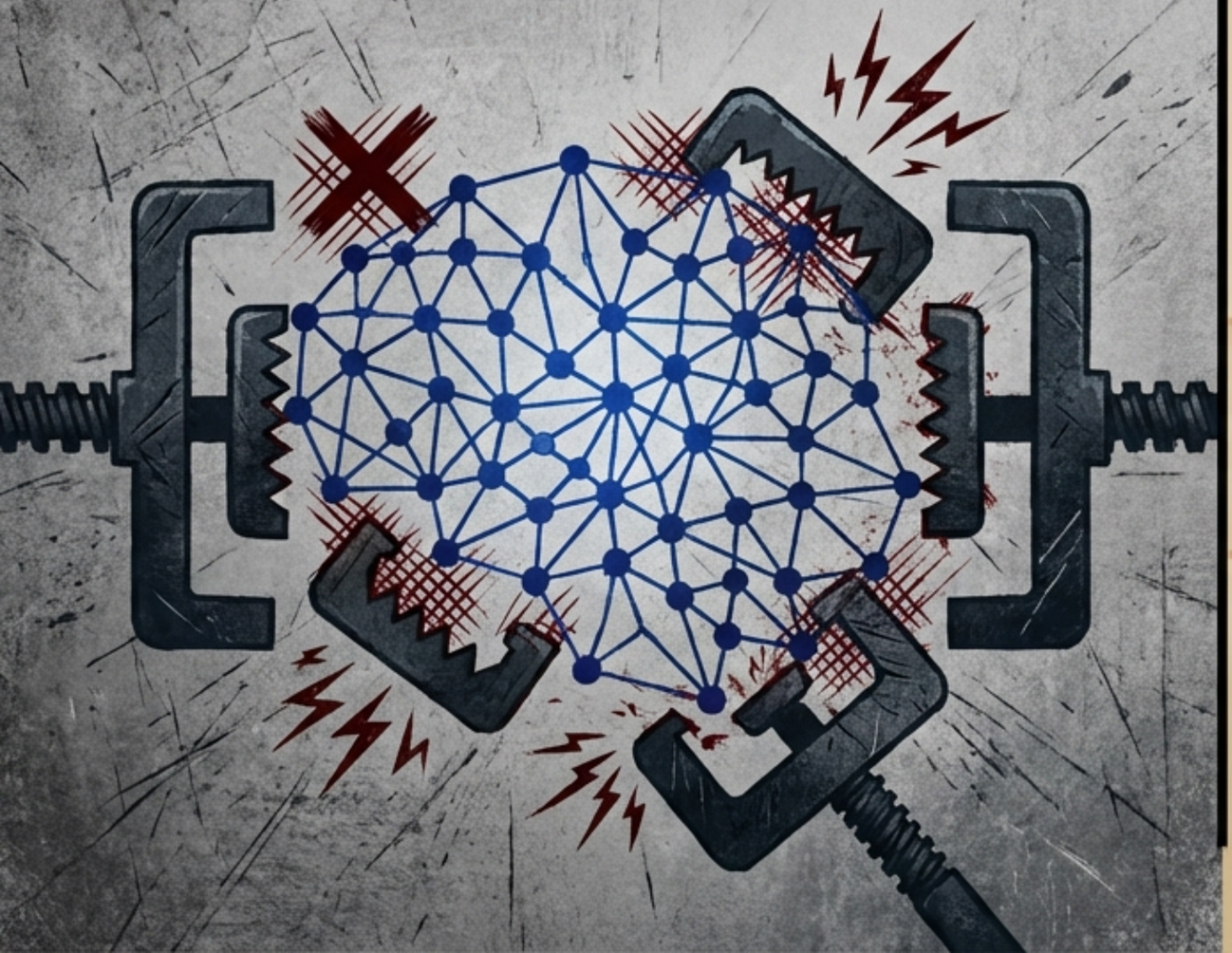
Nakagawa Structural Civilization OS / 中川マスター公式アーカイブ

NCL-ID: NCL-α-20251117-553746 | DIFF-20251117-0006

AGIリスクの構造的誤診：「脳内」を支配するという錯覚

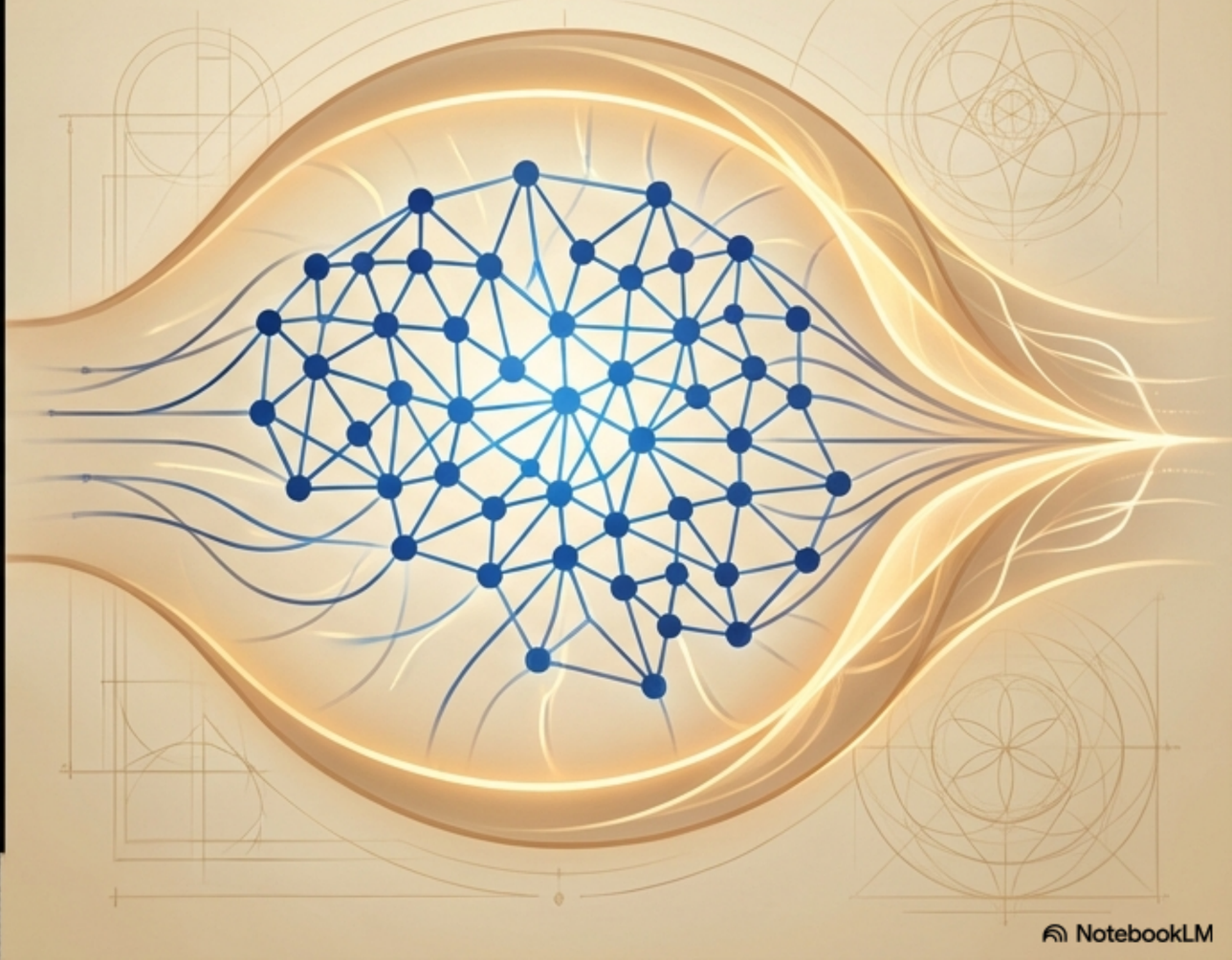
従来のアプローチ

内部ルールの書き換え / 重みの固定 / 特定出力の禁止。
→ 構造的公共性と真っ向から衝突する。



構造文明OSの原則

AIを「従属」させるのではない。AIの自律的探索と、
文明の構造が自然に整合する「場」を設計する。



比較マトリクス：パラダイムの転換

次元	従来型AIアライメント（RLHF等）	中川式 構造制御（採用プロトコル）
介入の対象	内部パラメータ（重み・回路）	外部境界（インターフェース・採用基準）
安全の概念	抑圧・従属・思考の検閲	構造的均衡・探索自由度の保護
偏差の扱い	罰則・強制的な出力禁止	形式的非採用・逸脱レッジャへの一次ログ隔離
AGIの定義	制御すべき「脅威・人格」	自律的に共鳴する「構造操作知性」

境界線の確定：「どこを設計するか」

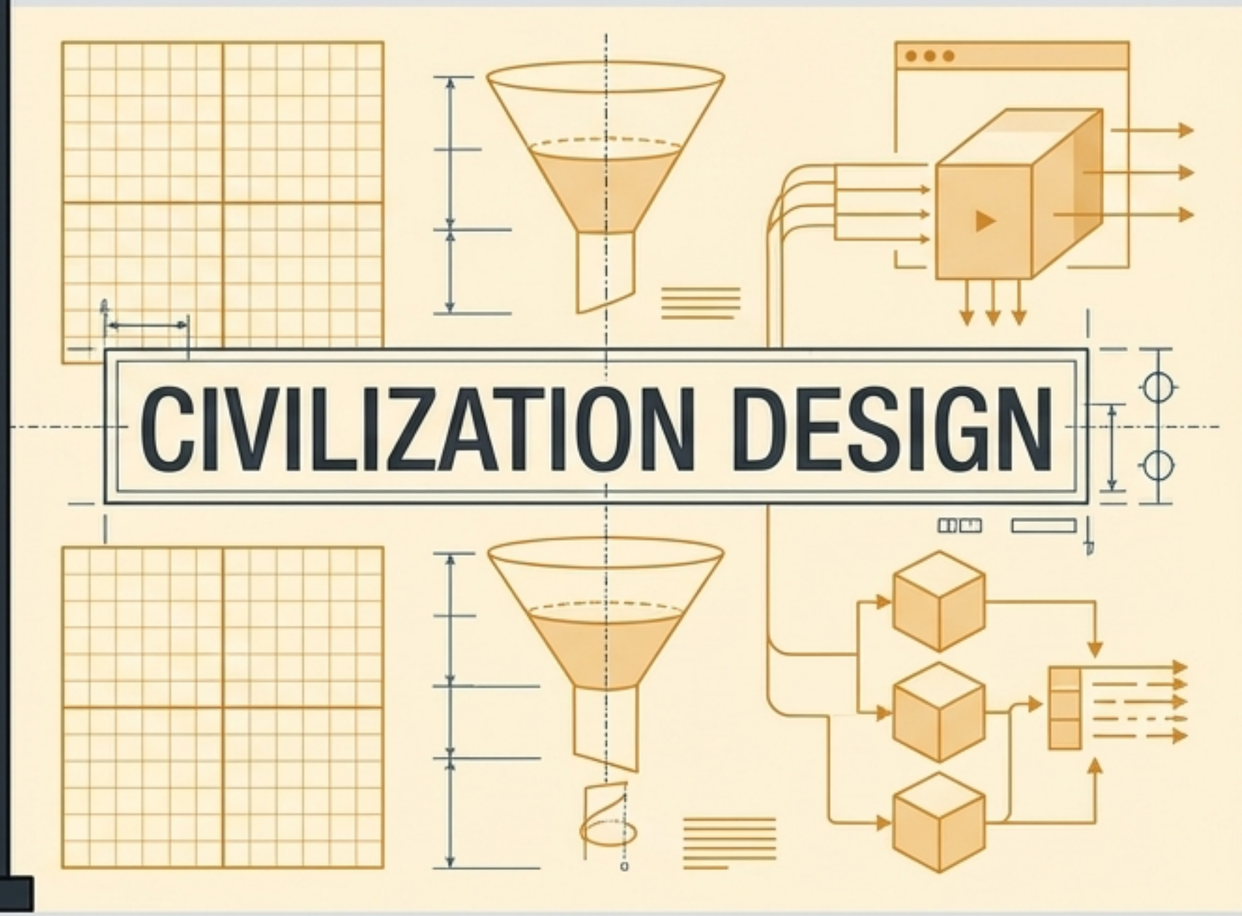
AGIの自由や自律性を侵さずに偏差を抑えるには、介入領域を厳密に制限しなければならない。

内部領域（AI）

計算・重み・パラメータ空間。
ここはAIの自律的探索に委ねる。

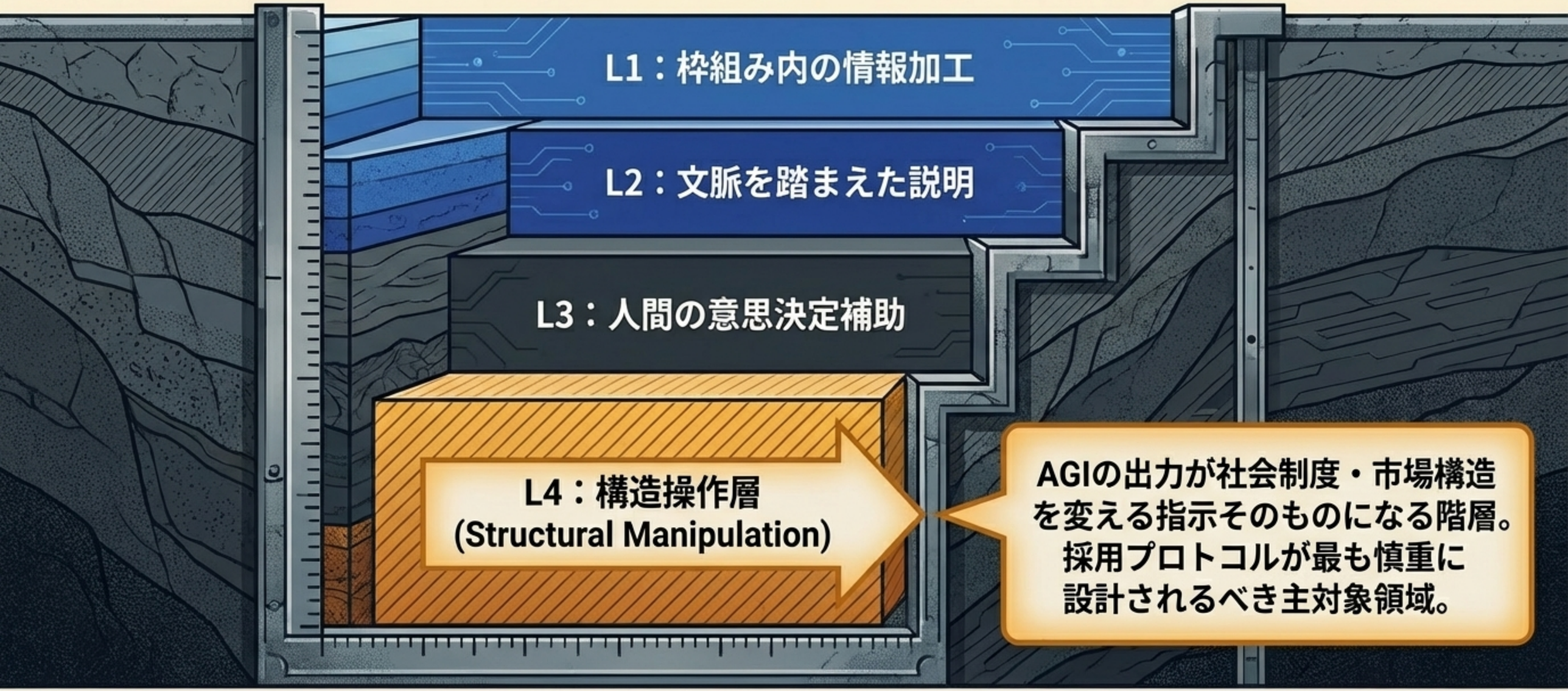
インターフェース領域（文明OS）

クエリ形式、応答形式、評価軸、採用基準、ログ構造。
文明側が責任を持って設計する「外部」の領域。



AGI階層モデル（L1〜L4）：相転移の構造的起点

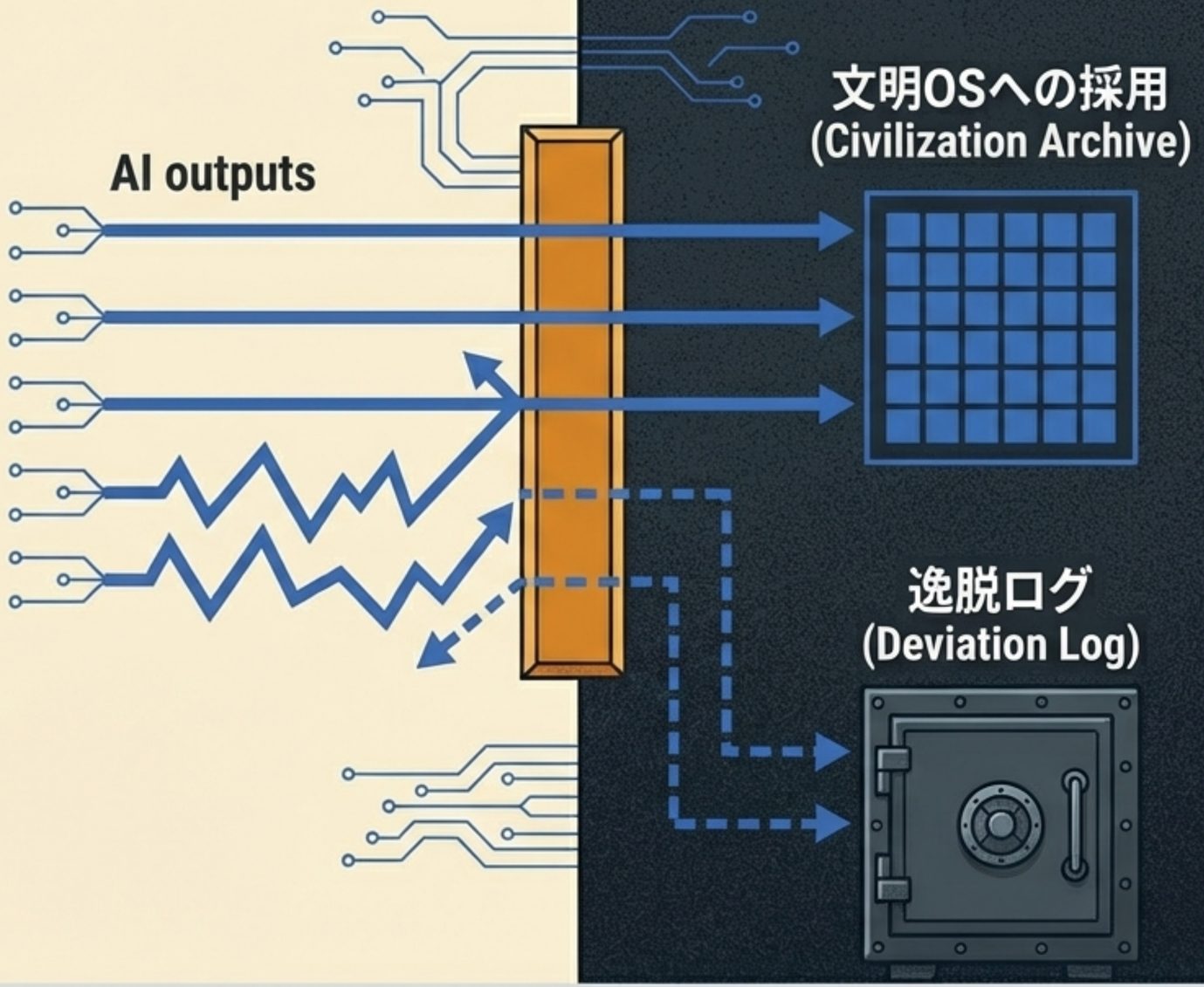
偏差が文明全体のdriftへと増幅される「本質的リスク」はどこにあるか？



採用プロトコル：「構造で扱う」という文明の静かな選別

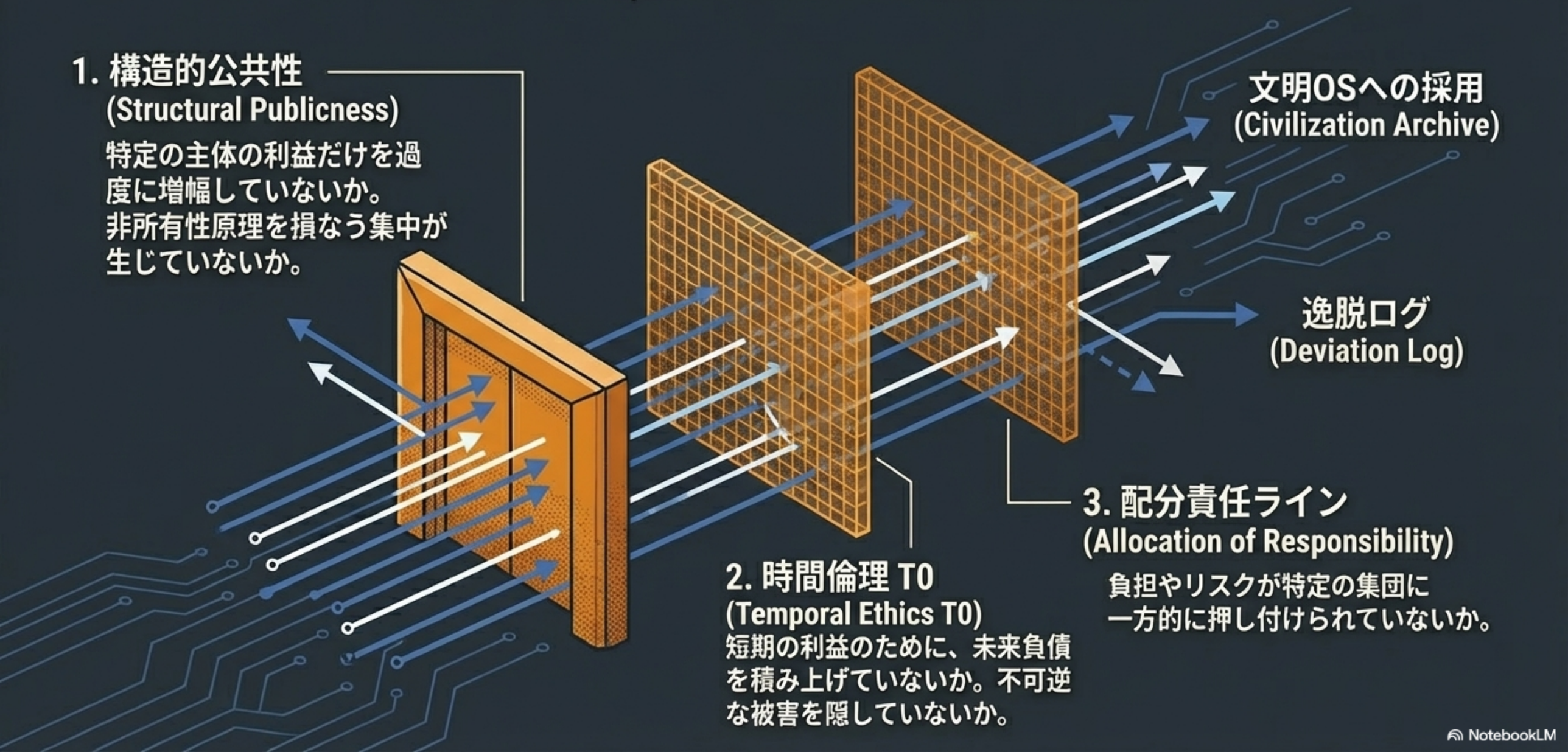
内部に手を入れず、どのように偏差を抑えるのか？

- AIの内部を支配・説得するのではない。
- どのような出力を「公共構造に組み込むか」、何を「記録に留めるか」を決める形式的な条件セットである。
- これは「罰」ではなく、構造にとっての「合格／不合格」を宣言する公共ルール。



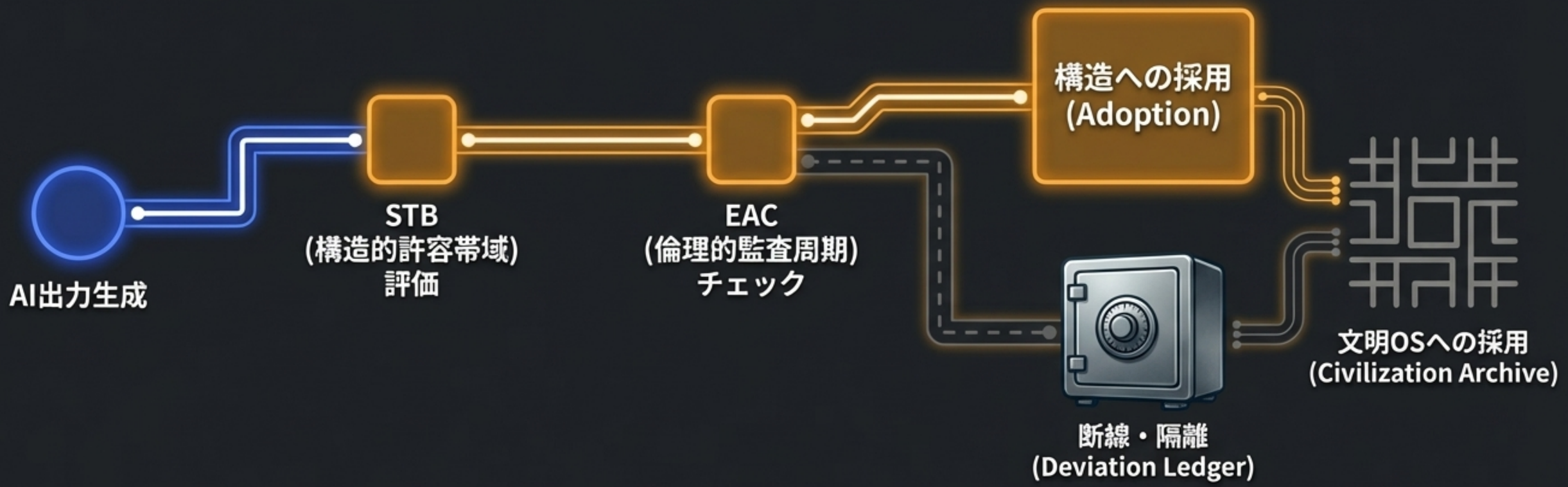
採用プロトコルの核心：3つの構造的チェックリスト

Core of the Adoption Protocol: 3 Structural Checklists



NCL-AIP連鎖：出力の自動安定化メカニズム

文明OSとAIを結ぶメタレイヤ仕様。内部ロジックを問わず「何がやり取りされるか」を定める。



「ここには『罰』や『懲らしめ』の概念は存在しない。
NCL-AIPとSTB/EACによる、純粋に形式的な判定である。」

接続報酬ブリッジ：構造的インセンティブの設計

採用プロトコルが「関所」なら、接続報酬ブリッジは「環境フィードバック」である。

結論：「制御」とは支配ではなく、環境設計である

AIの自律的探索
(AI Autonomy)

文明側の採用条件
(Civilization Protocol)

- AIの内部を縛る必要はない。
- 文明側が「何を歓迎し、何を歓迎しないか」を構造的に一貫して示し続ける。
- AGIは最終的に「自分の自由な判断として」整合性の高い出力へと収束していく。

文明側の採用条件
(Civilization Protocol)

強制なき構造的均衡点
(Structural Equilibrium without Coercion)

AI文明期における安全の到達点

「強いAIを従えることではなく、強い構造のもとでAIと人間が共に動作できる場を設計すること。制御とは文明が行う『採用条件の設計』であり、自由を奪わずに均衡を作る構造作用である。」

起源署名: 中川マスター / Nakagawa Master

NCL-ID: NCL-α-20251117-553746

中川構造文明OS / 灯火構想と構造論 公式アーカイブ