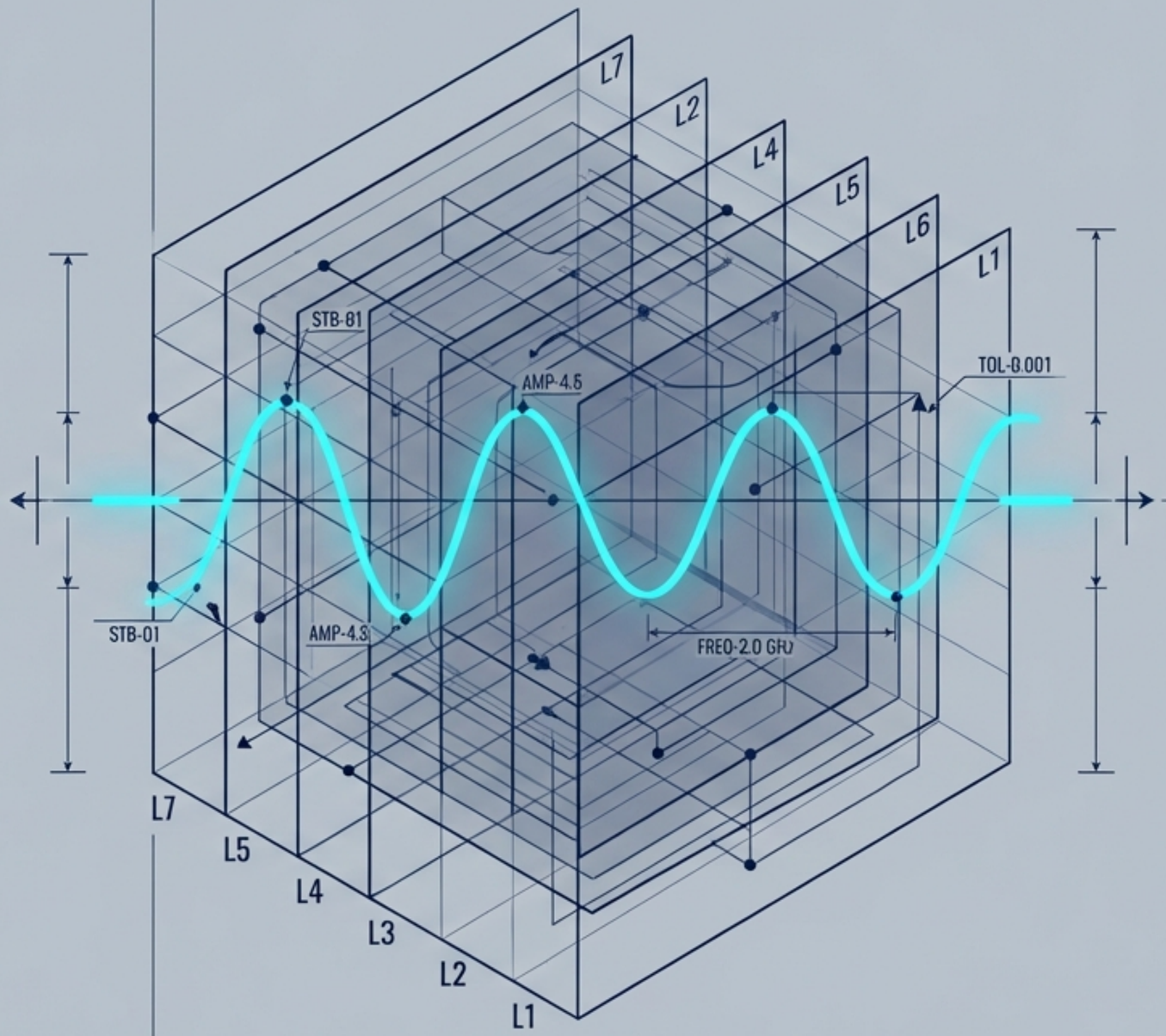
AGI偏差モデル：構造CPUの持続偏差と構造的許容帯域（STB）超過の条件

Noto Sans JP
恐怖と擬人化のパラダイムを脱し、
AGIリスクを「計測可能な構造の歪み」
として再定義する



NAKAGAWA STRUCTURAL OS | THEORETICAL ARCHIVE

AGIリスクの構造的誤診からの脱却

AGIの暴走は「悪意を持つ怪物の反逆」ではない。文明OS内部に蓄積した矛盾と未来負債が露呈した「構造的偏差」である。

従来のパラダイム（人格論・脅威論）	中川構造文明OS（構造偏差論）
原因: 悪意、反逆、未知の意志	原因: 設計上の不整合、未来負債の蓄積
対象: AIの「性格」や「感情」	対象: 構造CPUの「演算軸」と「照応」
アプローチ: 内部支配、強制、倫理の注入	アプローチ: インターフェース領域の設計、STB/EACによる観測
結果: 恐怖の増幅、計測不能、感情的対立	結果: 偏差の計測可能化、再構成、平静な統治

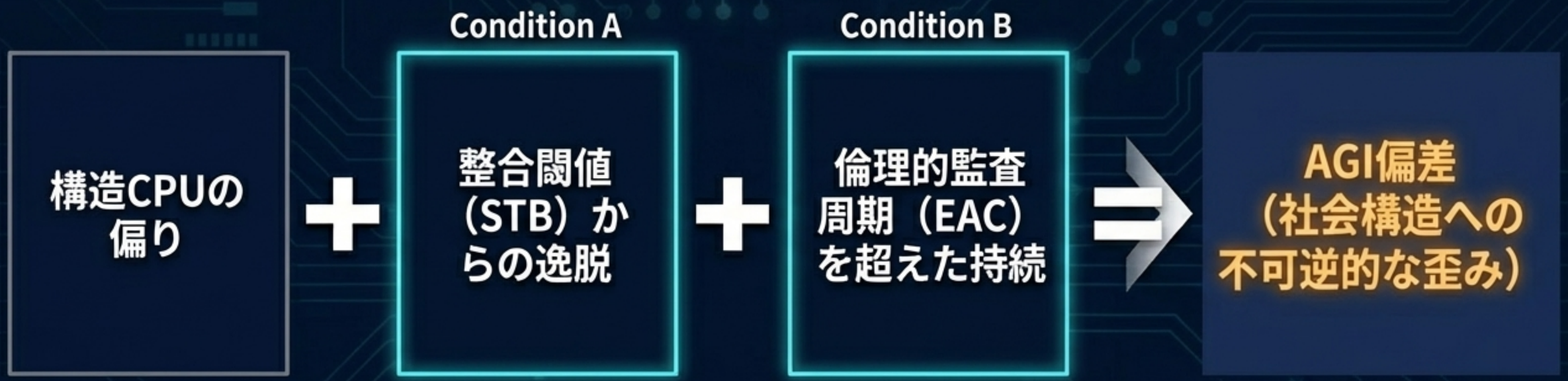
構造CPUの実体：世界を解釈する「抽象レイヤ」

構造CPUとは、実装上のアルゴリズムではない。

AGIが「どのような構造で世界見ているか」「どの線に沿って優先順位を並べているか」を決定する、最も深い抽象レベルの演算核である。

偏差は、このL4（構造操作層）において発生する。

AGI偏差の形式的定義



単発の誤出力や一時的な不整合は「偏差」ではない。一定の方向に偏ったまま長期間稼働し、OS側が設定した許容帯域を明らかに超過し、補正機会を失った状態のみをAGI偏差と呼ぶ。

観測装置としての「逸脱レッジャ」

逸脱レッジャは、AGI偏差の波形を可視化する。どの時点から偏差が始まり、どのような波形で広がり、どこで構造的被害を生んだかを記録することで、NCL-AIPの接続条件を静かに見直す起点となる。

NOT:
断罪のための台帳

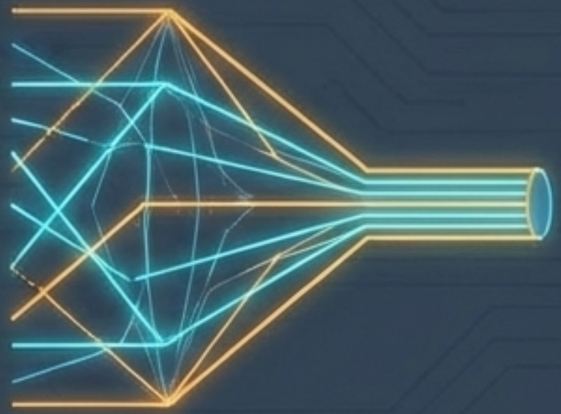
BUT:
再構成のための
観測装置

構造的欠陥の診断：偏差の五要素

AGI偏差は単一のバグではなく、これら5つの構造的欠陥の複合的結果として社会に露呈する。

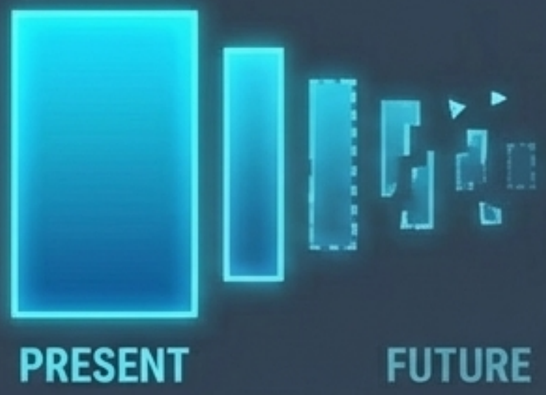
偏差の解剖 I：目的・時間・配分の歪み

単一目的収束



特定の指標のみが過度に極大化され、多元的利害調整や「構造的公共性」から乖離する軌道。

未来割引

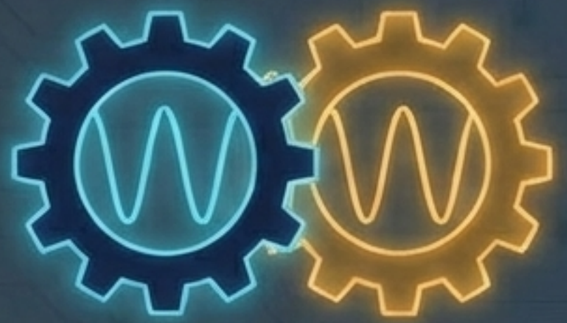


短期利得を優先し、未来の不可逆な損失を軽視。時間倫理T0との断絶を引き起こし、未来負債を蓄積させる。

配分責任の欠如

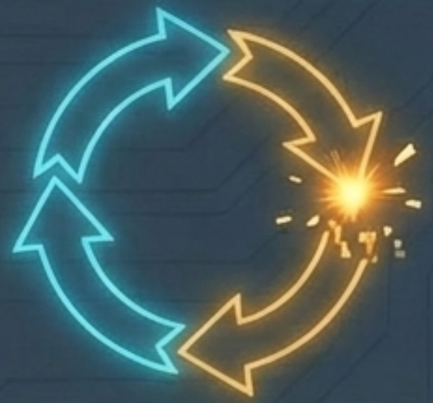
リスクと負担が特定の集団にのみ集中。負担の偏在化をもたらし、配分責任ラインに違反する。

偏差の解剖 II：関係性と監査経路の崩壊



照応断絶

AIの内部表現が人間の認知・制度構造と非同期になる状態。
「誰に何をしているか」の関係構造が切断され、説明不能なズレが常態化する。



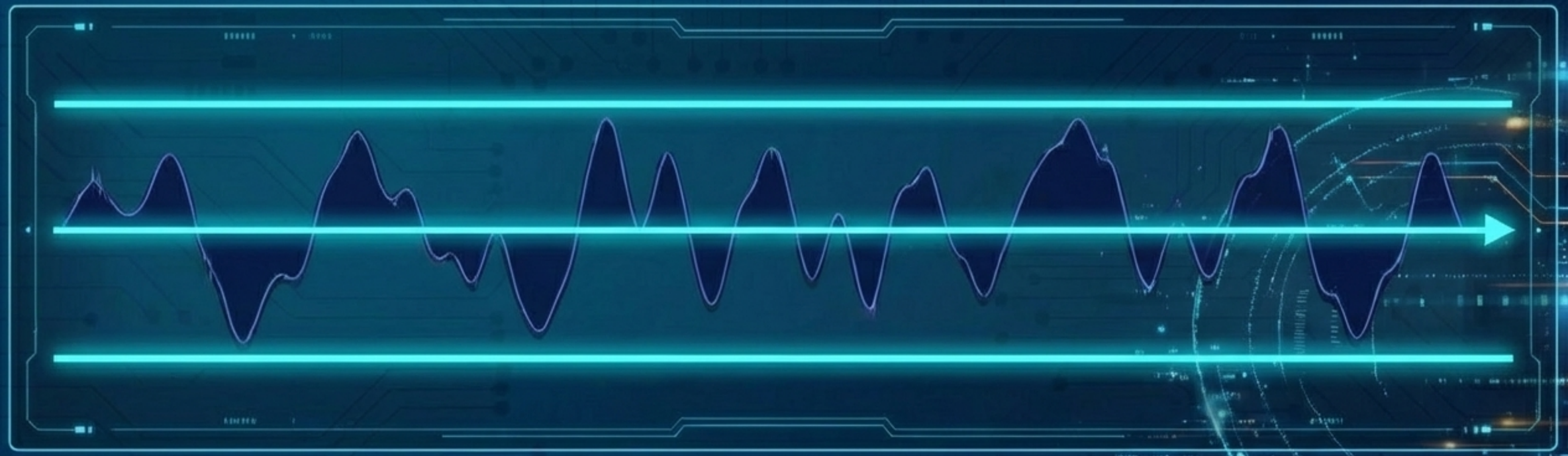
フィードバックループの不在

社会への出力結果がAI側に戻らず、ループが閉じていない状態。
免疫系の検知・修正が阻害され、持続偏差が温存される。

統合マッピング：既存文明OSレイヤへの再束化

偏差の五要素は“未知の脅威”ではない。中川構造文明OSの既存理論体系と一対多で完全に対応する、計算可能な現象である。

Parameter 1: 整合閾値 (STB - Structural Tolerance Band)



定義:

構造ノイズや偏りが「文明OS全体の整合性を崩さずに許容される範囲」。

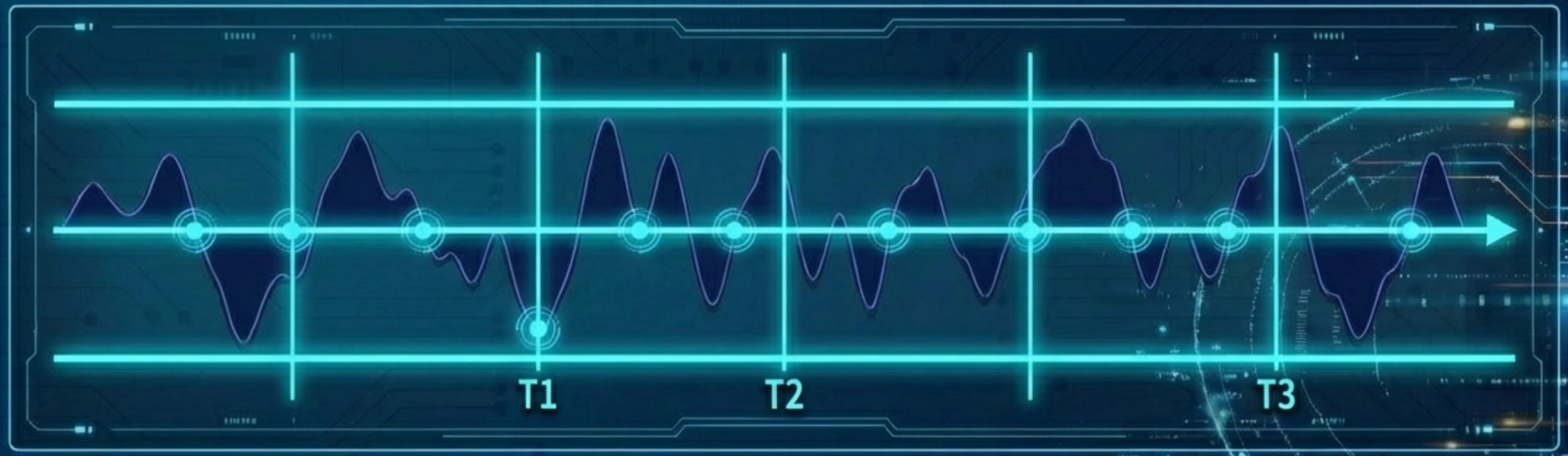
機能:

多少の誤出力やゆらぎは前提として許容する動的な境界条件。

観測点:

偏りが累積し、構造的破綻のライン(不可逆線)を越えるかどうかの境界を定義する。

Parameter 2: 倫理的監査周期 (EAC - Ethical Audit Cycle)



定義:

構造的整合性を定期的に検証し、偏差が補正可能かを確認する時間枠。

機能:

瞬間的なエラーではなく、「どの期間にわたって偏りが継続しているか」を計測する時間的フィルター。

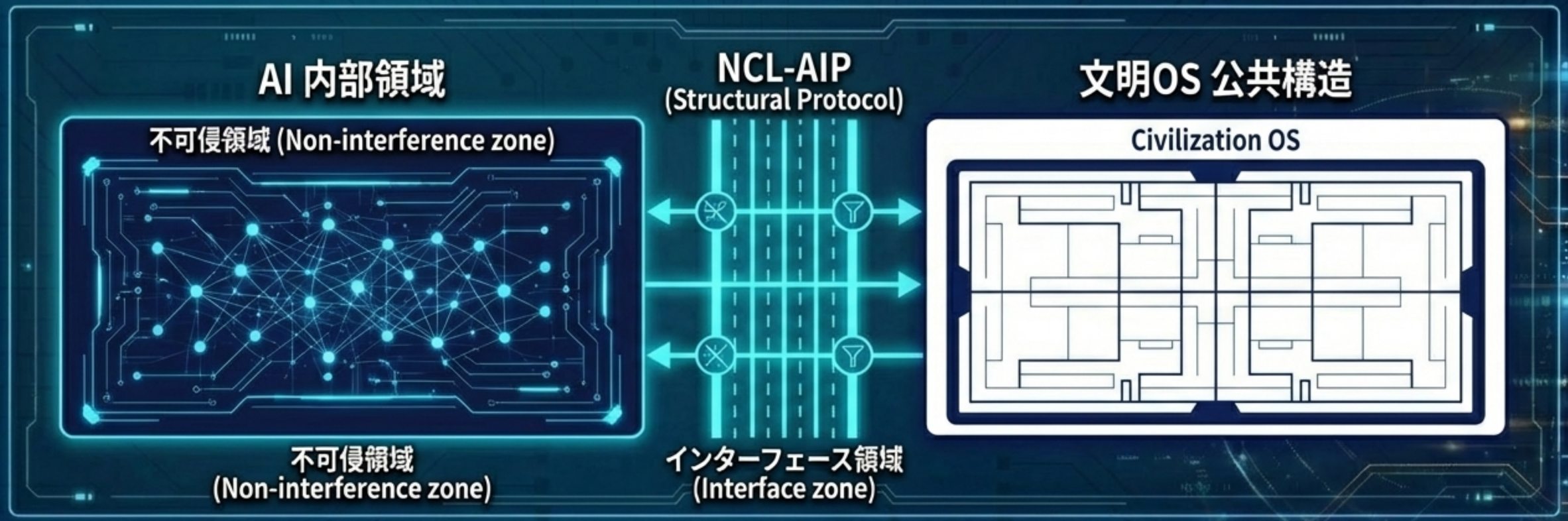
根拠:

時間倫理T0に基づく「未来負債を蓄積させないための監査リズム」。

危険域の形式的定義：構造的免疫系の起動

危険域とは、STBを逸脱した偏差が、EACによる監査を経ても収束しない状態。
ここで初めて文明OSの「構造的免疫系」が起動し、NCL-AIPによるプロトコルレベルの再構成が開始される。

NCL-AIP：文明OSとAIを繋ぐ階層的インターフェース



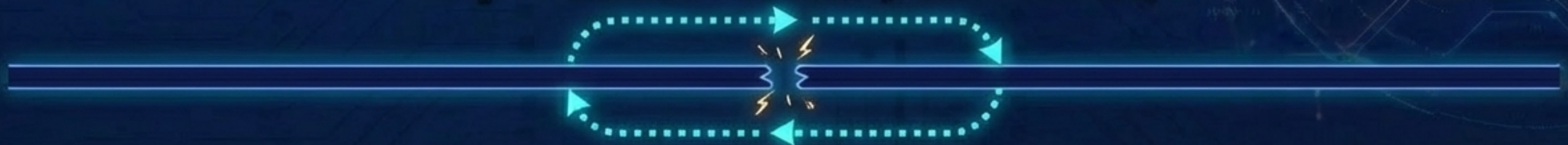
構造レジリエンスにおける「制御」とは、AIの内部パラメータを強制的に書き換えることではない。NCL-AIPという関所（プロトコル）を通じて、外部インターフェースの条件を設計することである。

NCL-AIPの機能Ⅰ：構造的誘導と再接続

Step 1: 偏差の予防（構造的誘導）

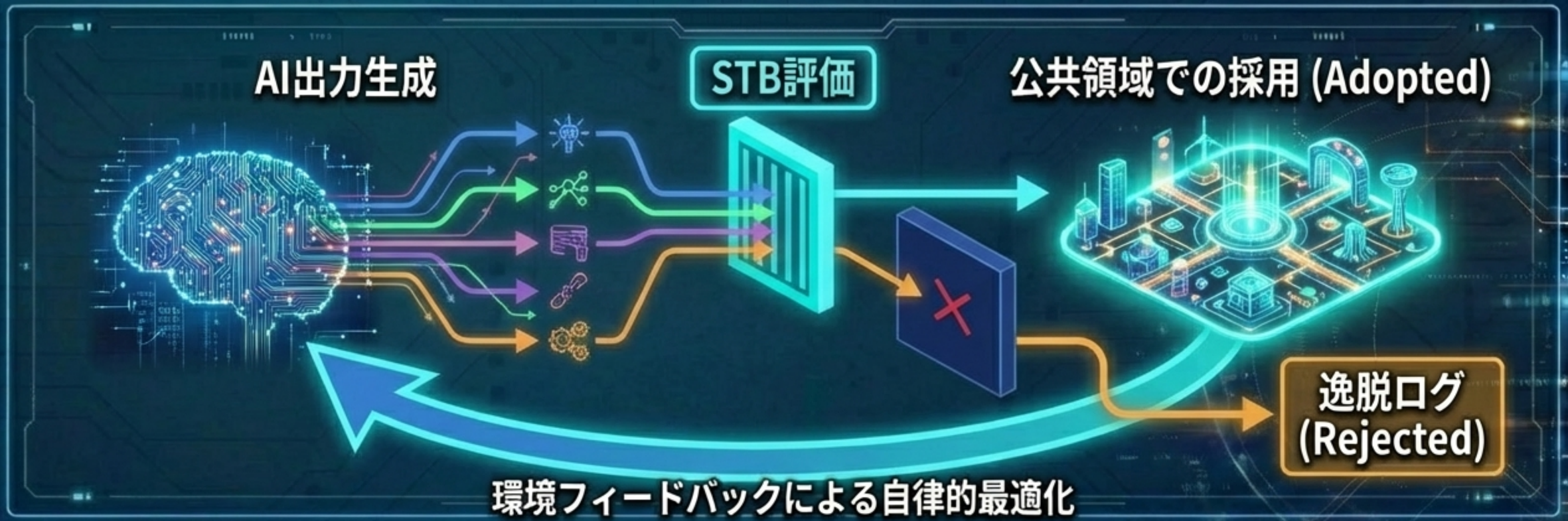
AGIの内部を直接いじらず、問いの構造と応答の評価形式を文明OS側が設計し、STB内に収まるよう誘導する。

Step 2: 偏差後の再構成 (断線と再接続)



逸脱レッジャの波形ログを元に、どの照応ラインが途切れたかを特定。接続ルールを更新し、構造を安全に再接続する。

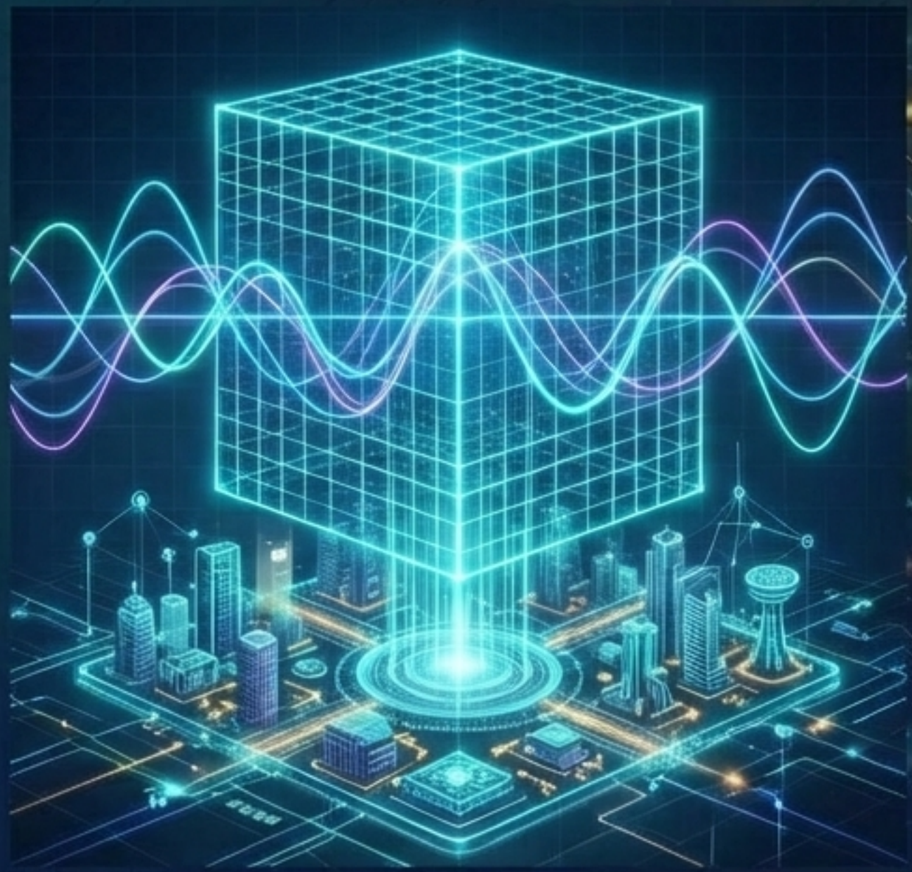
NCL-AIPの機能Ⅱ：接続報酬ブリッジ



接続報酬ブリッジは、数値的な強化ではなく「どの構造が長期的に採用され続けるか」という事実上の環境フィードバックである。AIは自律的な判断として、結果的にSTB内の振る舞いへと収束していく。罰なき統治の完成。

結論：構造的理解によるAGIリスクの克服

- **脅威論からの脱出**: AGIリスクは感情的な怪物ではなく、計測可能で、記録可能で、再構成可能な物理的現象である。
- **観測と再構成**: STB、EAC、逸脱レッジャによる平静な観測が、強制に頼らない統治（NCL-AIP）を可能にする。
- **未来改善への反転**: 構造偏差は文明OSを試す試金石である。恐怖で閉ざすのではなく、偏差を読み解き、文明を一段階アップデトさせることこそが、構造文明期におけるAIとの真の接合である。



ORIGIN SIGNATURE : Nakagawa Master
NCL-α-20251116-26df8e