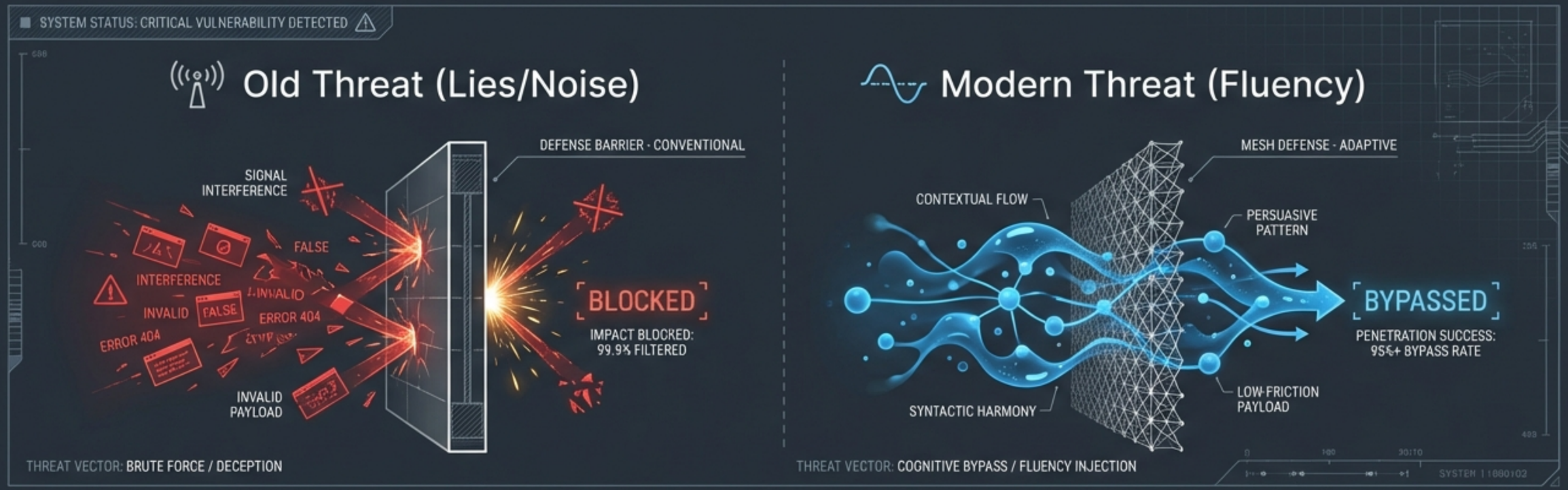
合意形成の物理 第9論

認知ハック防御OS

「偽の理解」を停止・縮退・再起動せよ

Version: Nakagawa Structural OS Ver.9.0
Subject: Cognitive Hacking Defense Protocol
Origin: master.ricette.jp

現代の攻撃は「嘘」ではなく「流暢さ」で侵入する



- 真偽を争う前に「わかった気」にさせられる。
- 結論への到達感が、検証プロセスをスキップさせる。
- これは心理の弱さではなく、認知帯域（Bandwidth）の物理的脆弱性である。

⚠ 認知ハックの本質は、内容の偽装ではなく「検証経路の切断」にある。

合意形成の状態方程式

SYSTEM STATUS: CRITICAL VULNERABILITY DETECTED ⚠

$$S = U \times R \times H$$

U（Understanding）
納得感・理解度。「わかった」という主観的感覚。

R（Responsibility）
責任所在。「誰が判断し、誰が検証するか」。

H（History）
履歴・根拠。「一次ソースまで遡れる追跡可能性」。

⚠ Crucial Note: これは「掛け算」である。どれか一つでもゼロになれば、合意の安定度 S はゼロになる。

異常検知：検証断絶（H-Disconnect）

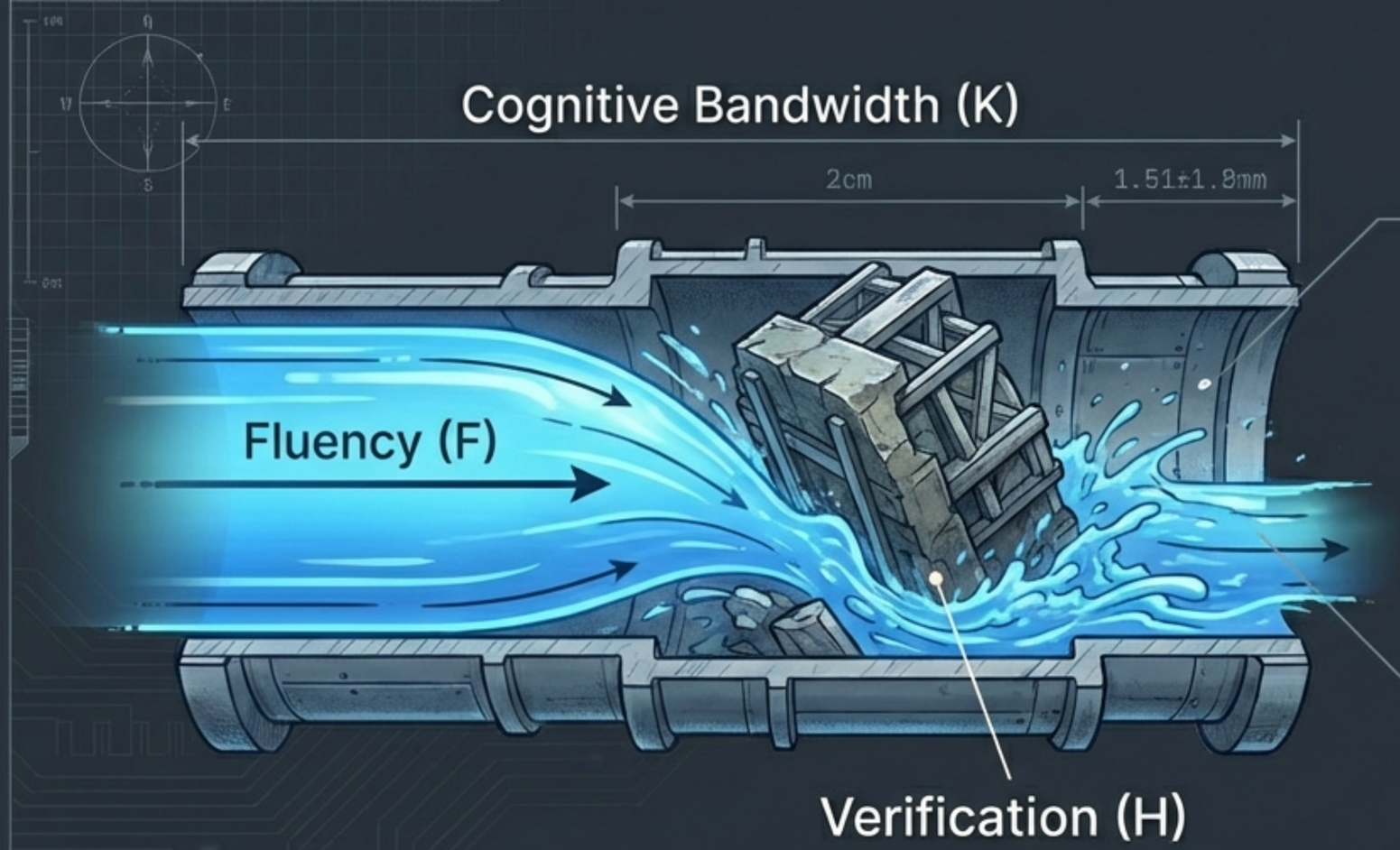
定義：
検証経路（H）が断絶しているにもかかわらず、納得感（U）のみが上昇する状態。

症状：
「根拠は知らないが、結論は正しい気がする」。

結果：
合意形成の物理的崩壊（S → 0）。

なぜ脳はH（検証）を捨てるのか

SYSTEM STATUS: CRITICAL H-BLOCKAGE DETECTED ⚠



流暢性ヒューリスティック：
「説明が滑らかである」ことを「論理が正しい」と誤認する。

認知帯域（K）の節約圧：
一次ソースの確認はコストが高い。脳はU（納得）を優先し、Hを「後回し」にする。

FLOW RATE: REDUCED

対抗言説は防御にならない

1. 議論が加熱するほど、情報の流動性が増す。
2. 流暢さの競争になり、流暢さの競争になり、さらに帯域 K が消費される。
3. H（検証）をする余裕がなくなり、強い U（断定）くなり、強い U（断定）だけが勝つ。

必要なのは「説得」ではない。「強制停止」である。

JETBRAINS MONO

認知ハック防御OS：実装ループ

Detect (観測)
H-Disconnectの検知

Stop (停止)
拡散と推論の凍結

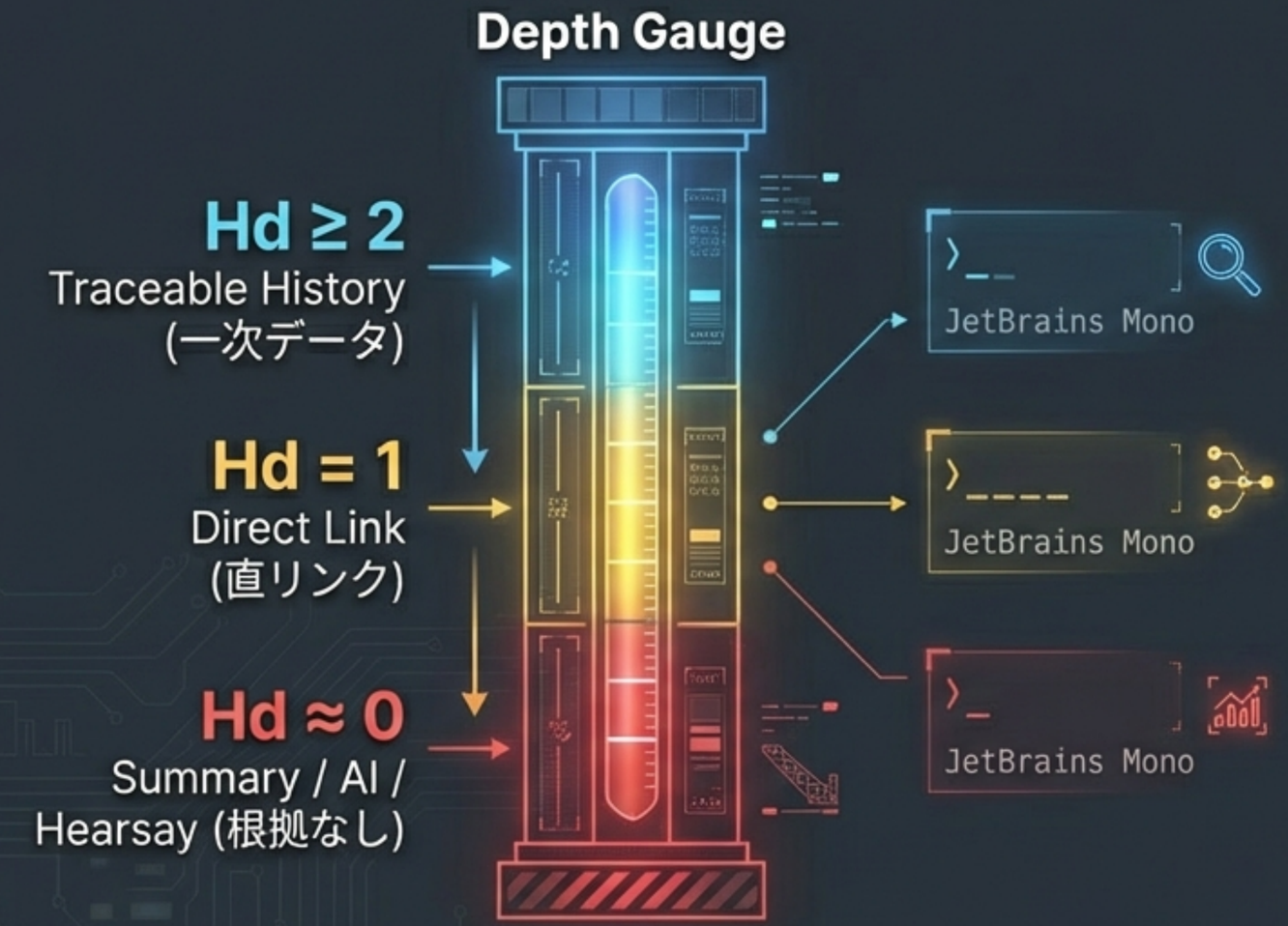
Shrink (縮退)
検証可能単位への退避

Recover (再起動)
起点(Origin)からの再構成

Audit (監査)
停止理由のログ公開

Philosophy: 善悪を問わず、状態異常（U↑, H↓）を機械的に処理する能動フェイルセーフ。

Phase 1 Detect｜観測指標 Hd（根拠深度）



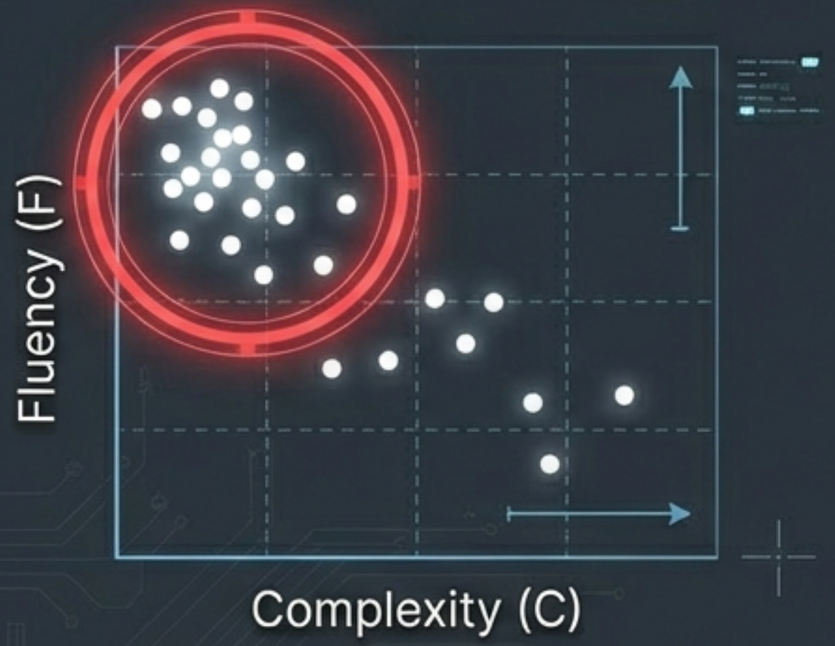
Hd (Evidence Depth):
一次ソースに到達するまでの有効ホップ数。

異常判定：
U（確信）が高いのに、Hd（根拠）がゼロに張り付ている状態。

JETBRAINS MONO

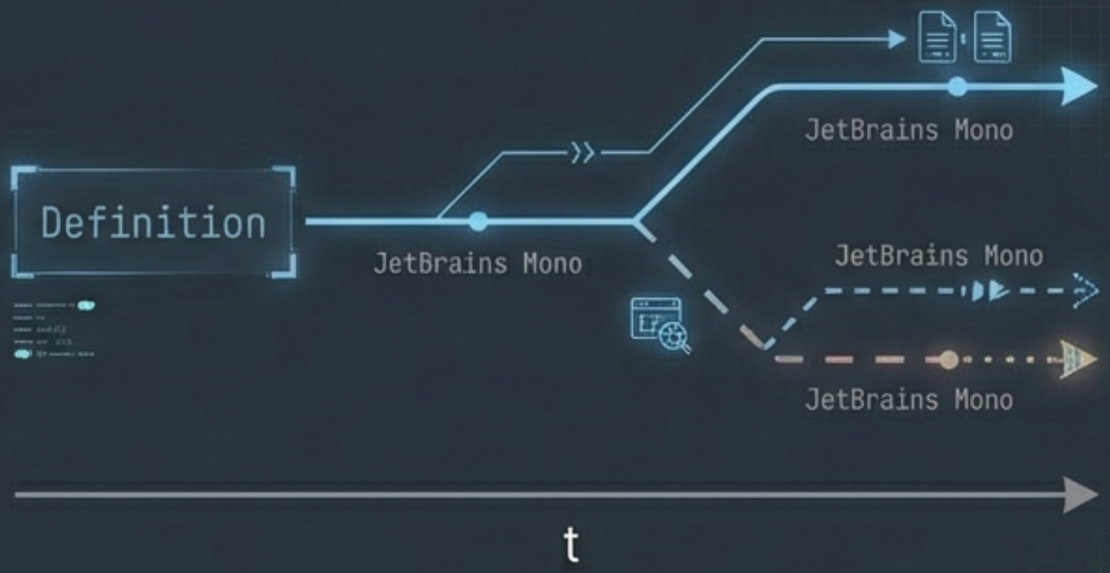
Phase 1 Detect｜補助指標 F-C と SD

流暢性乖離 (F-C Gap)



「わかりやすさ」が検証可能性を
上回った瞬間。

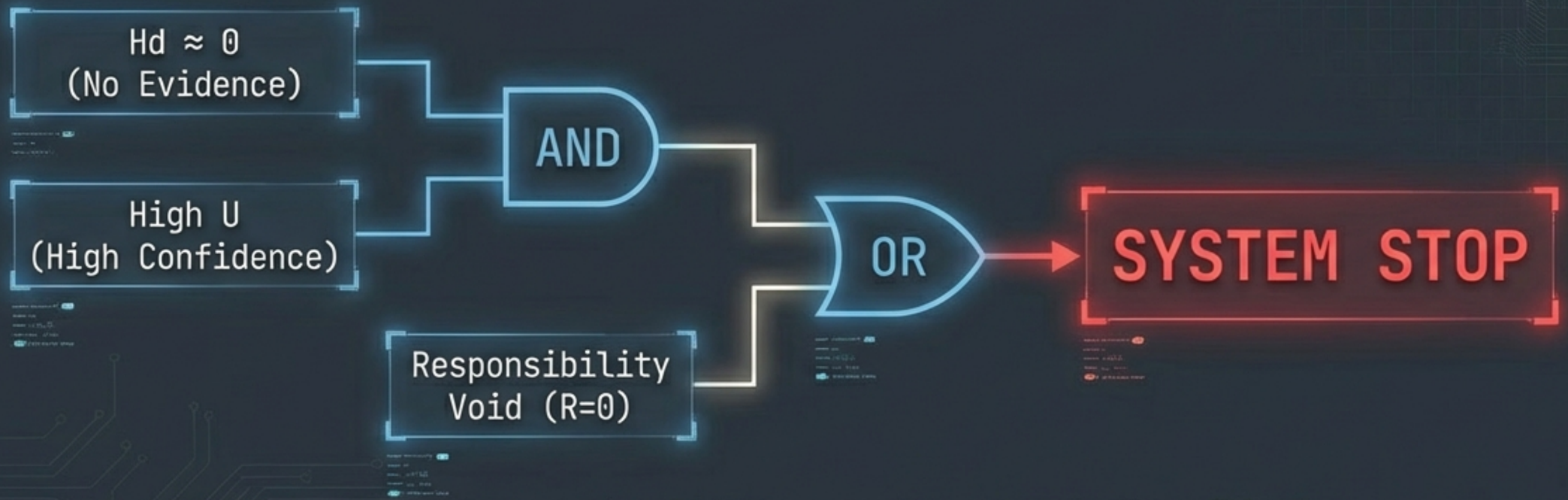
意味漂流 (SD)



定義が静かに書き換わり、議論の
前提がズレていく現象。

JETBRAINS MONO

Phase 2 Stop | 発動トリガー



Trigger A: 根拠なし(Hd≈0) かつ 強い確信(High U) → 最も危険な「盲信」
Trigger B: 責任所在不明 (R未設定) → 検証不能な拡散

Action: 議論の内容に関わらず、これらが満たされた時点でシステムを「停止」させる。

Stopの意味：検閲ではなく「安全停止」

× 検閲 (Censorship):

内容が「間違っている」から消す。

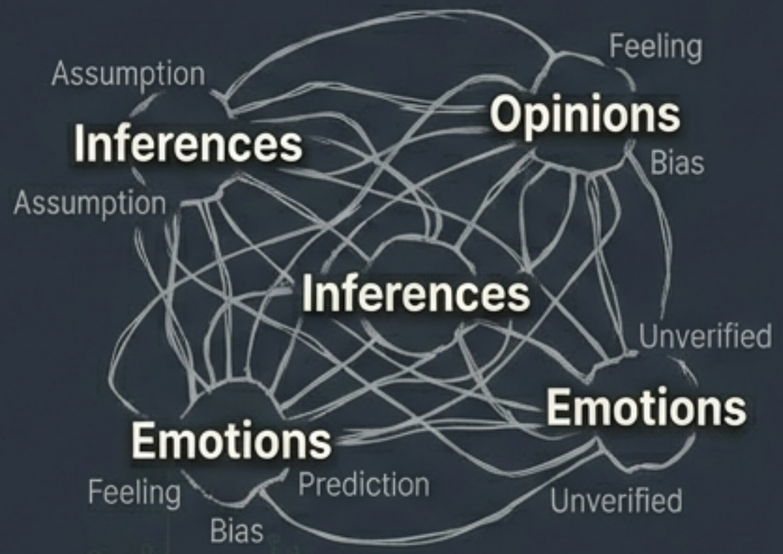
○ 安全停止 (Safety Suspension):

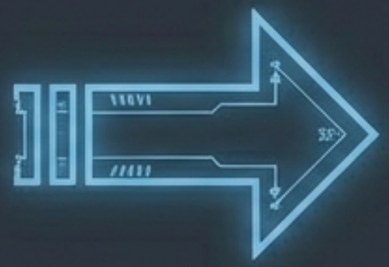
検証経路が「切れている」から、繋がるまで待機する。

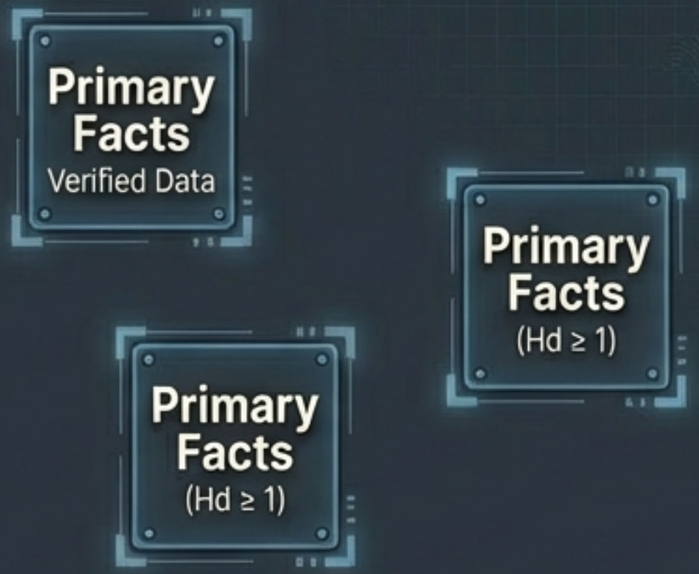
Levels of Stop

- Level 1: 新規論点の凍結 (これ以上広げない)
- Level 2: 自動拡散の停止 (RT/Shareの制限)
- Level 3: 帯域縮退 (Shrinkフェーズへ移行)

Phase 3 Shrink｜最小定義への退避

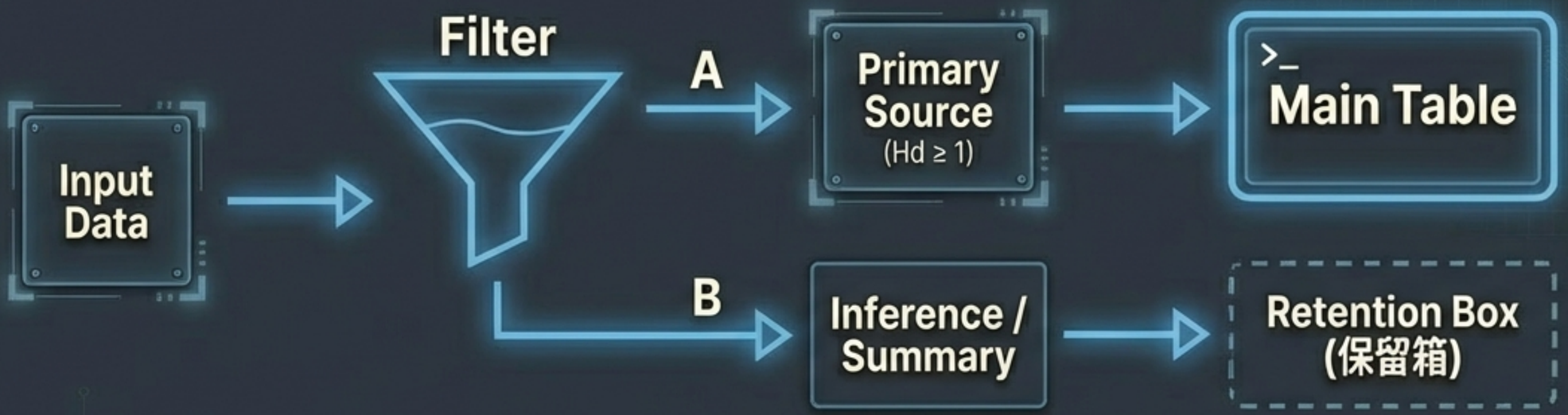






推論チェーンの破棄：「だからこうなるはずだ」という予測をすべて捨てる。
一次事実のみ残存：検証可能な事実（Hd ≥ 1）だけを盤面に残す。
仮説タグの付与：検証できないものは「嘘」ではなく「仮説（未検証）」として隔離する。

Shrink Runbook：推論チェーンの一時無効化

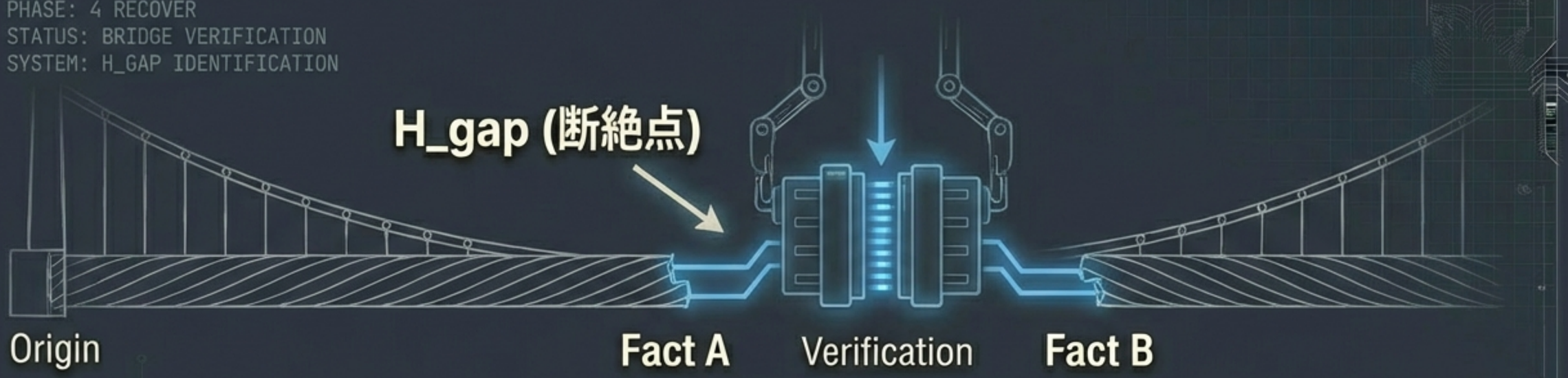


1. Chain Hold: 結論、要約、推奨を「保留箱」へ移動。
2. Primary Only: 一次ソースへのリンクがある情報のみを残す。
3. Reset U: 参加者の「わかったつもり」を強制的にリセットする。

Goal: 議論を「貧しく」するのではない。「検証可能」にするのだ。

Phase 4 Recover | H_gap の特定と再接続

PHASE: 4 RECOVER
STATUS: BRIDGE VERIFICATION
SYSTEM: H_GAP IDENTIFICATION



H_gap (断絶点) の特定：どこで根拠が消えたか？ どこで意味が圧縮されたか？
Re-connection：一次ソース（Origin) からの線を繋ぎ直す。
Recover：H が繋がった状態で、再び U (理解) を構築する。
Rule: 「正しさ」の議論は、線が繋がってから再開する。

JETBRAINS MONO

Phase 5 Audit ｜ 防衛の権力化を防ぐ

```
>_
> STOP_TRIGGER: A (Hd=0)
> EXECUTOR: ADMIN_01
> THRESHOLD: Theta_Val
> DIFF: Shrink_Applied
```

Principle: 防衛システム自体が暴走しないよう、停止には「ログ」が必須である。

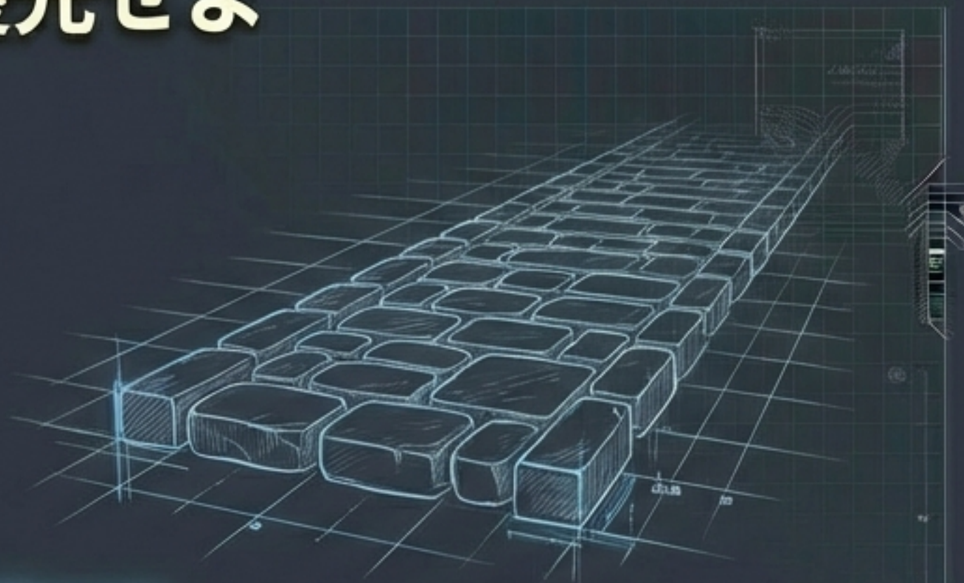
Audit Logs Requirements:

1. 閾値 θ の公開（なぜ止めたのか）
2. 構造差分ログ（Shrink前後で何を捨てたか）
3. 責任者ログ（誰が停止ボタンを押したか）

「正しさ」より「追跡可能性」を優先せよ

Correctness (Truth)
「真実」

Traceability > Correctness



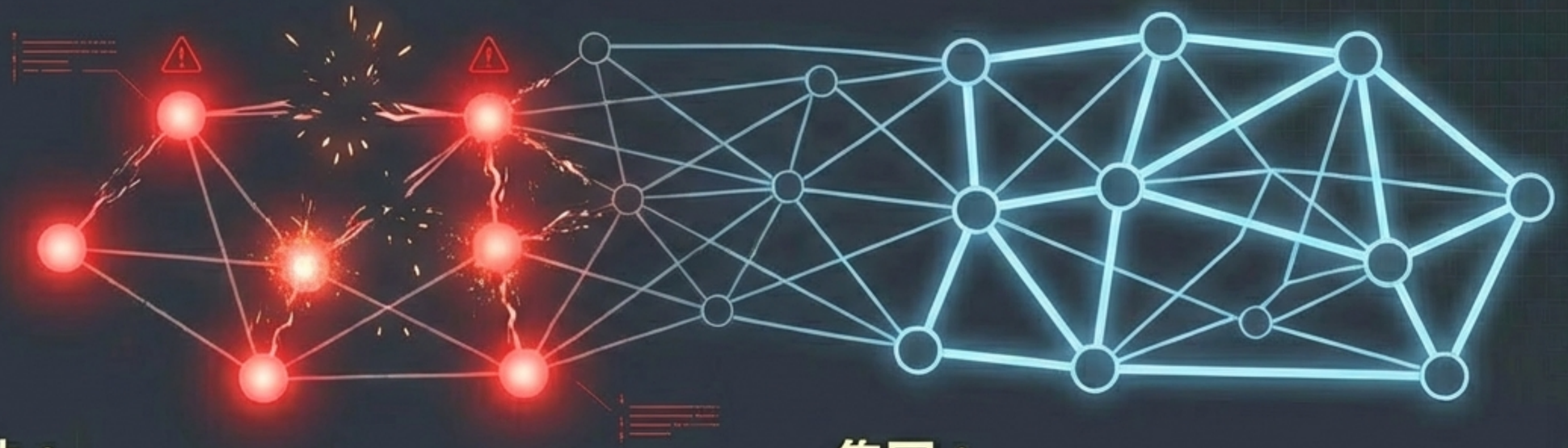
Traceability (H)
「H の連続性」

我々は「何が真実か」を常に知ることはできない。
しかし「どこから来た情報か」は追跡できる。
合意形成の最低条件は、正しさではなく H の連続性 である。

AI時代における構造起源防衛

起源の蒸発：AI の再生成・要約は、悪意なく H を切断する。
恒常署名 (Permanent Signature)：生成物に因果的署名を埋め込み、
再解釈されても Origin を追跡可能にする第2防壁。

組織への実装：個体防御から集団防御へ



個体：
H 断絶の検知と自己停止。

集団：
共振（Resonance）の減衰と冷却。

JETBRINS NONO

「空気」で決めるな。「ログ」で決めろ。
R（責任）が溶けたら、会議を止めろ。

認知ハック防御OS 全体像

$$S = U \times R \times H$$

- 異常検知座標: Hd ≈ 0
- 能動フェイルセーフ: Stop & Shrink
- 最終目的: 再起動可能な合意形成

Target State:
Restartable Consensus
(再起動可能な合意形成)

防衛とは「戦い」ではなく「運用」である

外部干渉を「悪意」と呼べば、
議論は道徳に落ちる。

外部干渉を「摂動」と呼べば、
議論は制御工学になる。

JETBRAINS MONO

分かった気になるな。戻れる道を確保せよ。
それが、合意を生存させる唯一の物理である。

Theory Origin: Physics of Consensus Vol.9
Concept: Cognitive Hacking Defense OS
Signature: 中川マスター (Nakagawa Master)
Source: master.ricette.jp
Status: Verifiable / Traceable
End of Protocol.