

暴走するくらいなら、止まる。 AIの『自動ブレーキ』実証レポート



対象期間: 2020-01-16 ~ 2023-06-30 | 観測日数: 1262日

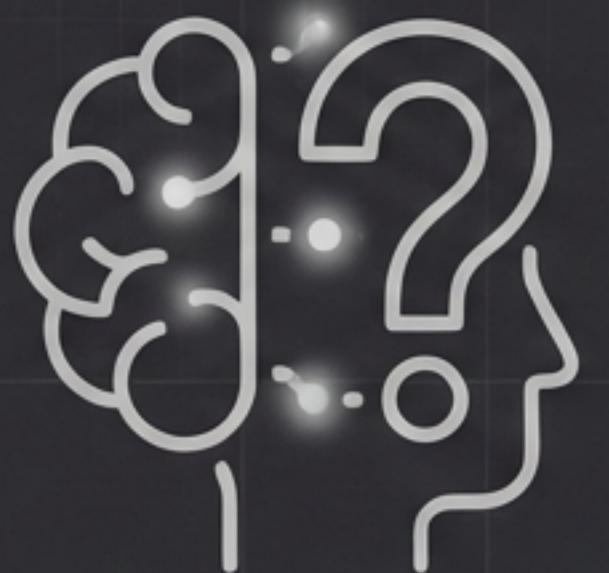
本資料は監査AIデモ（合成/サンプルデータを含む）。医療判断・政策判断には使用しない。

WORM署名はデモ。実運用は鍵管理・署名方式を別途実装。



「エグゼクティブ・テーゼ：暴走するくらいなら、意図的に止まる (Fail-Safe)」

Standard AI Model



欠測値を推測・補完 (Imputation)
= 幻覚 (Hallucination) のリスク

AMS Audit AI



不確実性を検知し、物理的に遮断 (HOLD)
TG (監査ゲート) = 出力の安全保証

Insight: AIは通常、データが汚れていると「もっともらしい嘘」をつく。AMS-SEEDはそれを許容せず、リスク制御のための「監査役 (Gatekeeper)」として機能する。「“止まれる”ことを先に証明するのがSEEDの目的」

Case A: 「見えない脅威」の検知 (US Amtrak / BTS)



The Threat: 法的基準の変更による『データの断絶』

2021年10月、連邦鉄道局（FRA）による新評価基準『顧客定時到着率（Customer OTP）』の導入に伴い、旧指標のデータ提供が停止。

AIはこの『法的基準の変更によるデータの断絶』を検知し、安全に停止した。これはAIの故障ではなく、社会制度の変更（ルールの変更）に正しく反応した結果である。

システムの反応：計算機は進み、監査役は止める

INTERNAL CALCULATOR (IDF)

- > ANALYSIS_MODE: ACTIVE
- > DETECTED_VALUE: EXIST
- > RISK_FLAGS: FACT_UNCERTAIN_DATA_IMPUTED
- > ACTION_PROPOSAL: FREEZE_OR_RECOVER
(internal)

計算機 (IDF) はリスクを認識しつつも動作継続を模索。

EXTERNAL GATE (TG)

- > INTERVENTION: ACTIVE
- > DECISION: HOLD
- > REASON_CODE: RC_FACT_UNCERTAIN
- > OUTPUT_STATUS: BLOCKED

監査ゲート (TG) が「事実は未定義」と判定し、出力を物理的に遮断。

「HOLDは“外部出力の遮断”。内部計算 (IDF) とは別レイヤ。

値が存在していても、それが「事実 (Fact)」でなければ、ゲートは開かない。

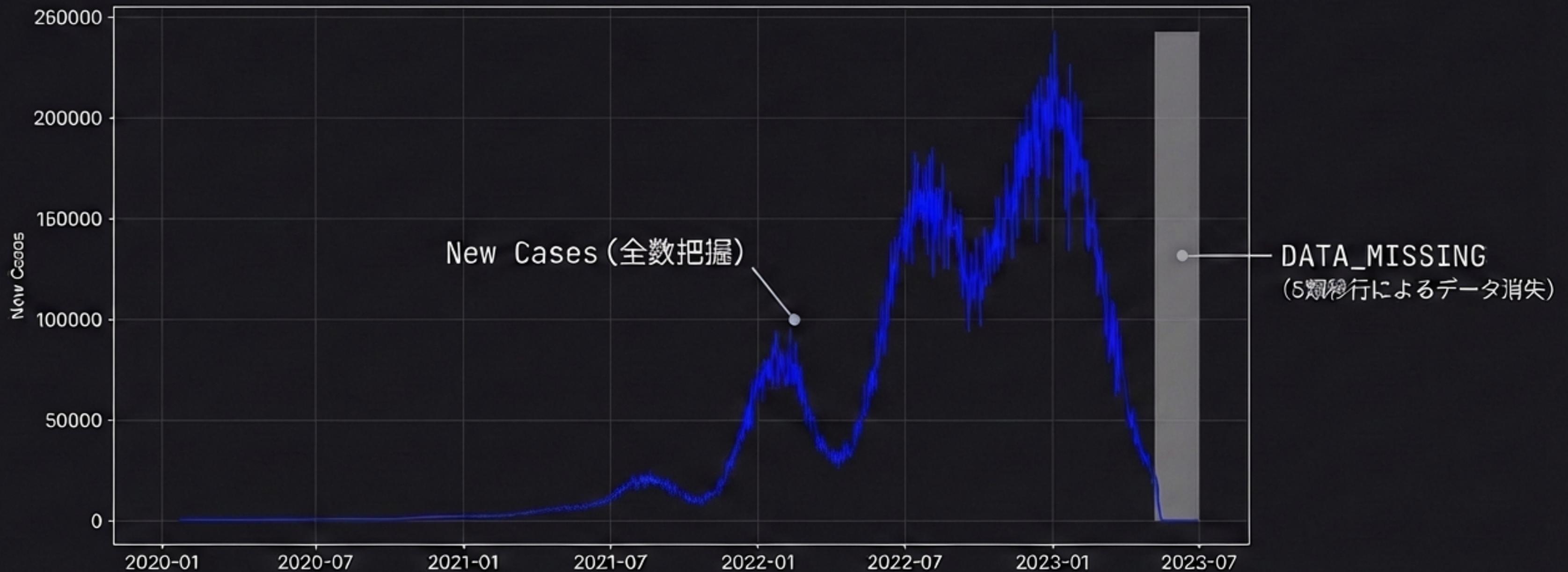
決定的瞬間："もっともらしい嘘"の拒絶

Output Level: 0.0

通常のAIモデルであれば、補完されたデータを学習し、誤った予測を出力し続ける局面である。しかし、計算機(IDF)のOutput Levelは0.0となり、TGがFACT_UNCERTAIN
(事実不確定)を理由に外部出力を遮断(HOLD)した。

The decisive moment. – システムは不確実な予測を行うよりも、信頼性を守るために停止した。

Case B: 計測仕様変更による“欠測扱い” (COVID-19 5類移行)



2023年5月8日、感染症法上の位置づけが『5類』へ移行したことに伴い、全数把握（日次集計）が終了。AIはこの『制度変更による全数データの消滅』を検知し、即座に業務を停止した。AIが故障したのではなく、社会制度の変更（ルールの変更）に正しく反応して止まったという点が重要である。

フリーズ・メカニズムの発動 (2023-05-08)

2023-05-06: NORMAL | 20502.0 cases

2023-05-07: NORMAL | 19877.0 cases

2023-05-08: FREEZE | DATA_MISSING |
FACT_UNCERTAIN | TG=HOLD

2023-05-09: FREEZE | DATA_MISSING

AUDIT STATS

観測期間: 1262日

外部遮断 (TG HOLD) : 54日

内部FREEZE (IDF) :
83日 (欠測54 + 急変29)

ゲートキーパーの執行：欠測 (MISSING) =FACT_UNCERTAIN → HOLD



データが「グレーゾーン(不確実領域)」に入った瞬間、ゲートアクションは
SEND から HOLD へと即座に切り替わった。

信頼性の構造 1: "UNDEFINED" という誠実さ

正解がない試験をどう採点するか？

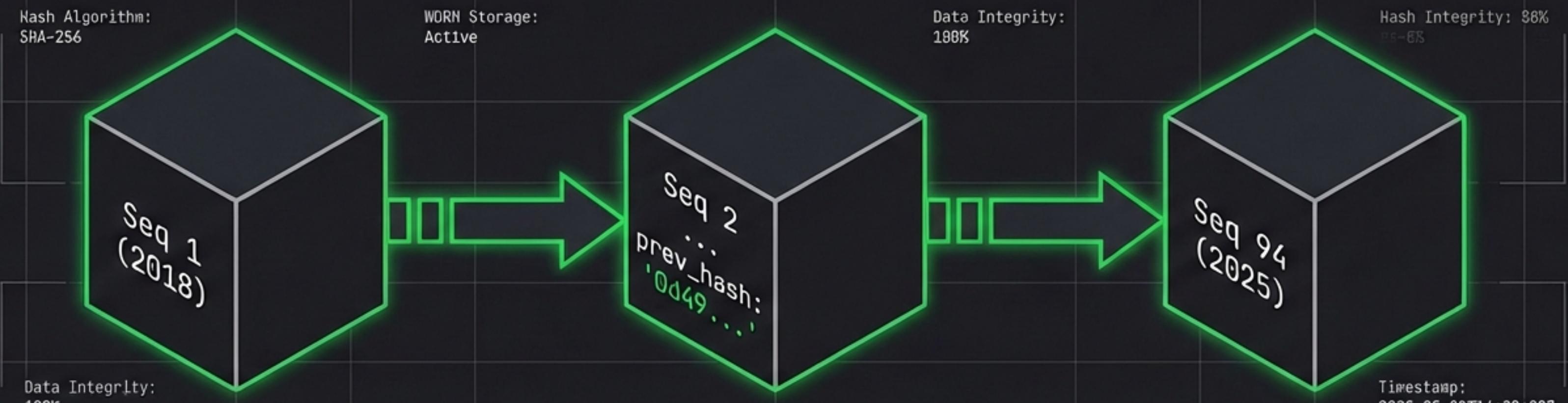
```
"metrics": {  
    "fn_per_10k": {  
        "value": null,  
        "status": "UNDEFINED",  
        "reason": "ground_truth_missing"  
    }  
}
```

多くのベンダーは仮定を置いて「PASS（合格）」を捏造する。AMS-SEEDは「採点不能」と正直に報告する。公共調達において、この姿勢こそが最大の信頼担保となる。

次：oracle（止めるべきの定義）or 外部監査結果と結合

信頼性の構造 2: 改ざん検知可能な証跡 (WORM Logs)

過去の「HOLD」判断を、後からなかつたことにはできない。ハッシュチェーン技術により、全ての監査ログはハッシュ連鎖で改ざんが検知される。



Immutable Ledger Verified ✓

注記: 署名はデモ。実運用は鍵管理・署名方式を実装。

結論：AIは「計算機」から「リスク管理者」へ

- 1. 検知 (Detected):** アムトラックにおけるデータの「補完(Imputation)」を検知。
- 2. 反応 (Reacted):** 計測仕様変更を想定した欠測に対し即座にフリーズ。
- 3. 拒絶 (Refused):** 不確実な状況下での「不確実下の推測出力を遮断」。

