Received: 23 December 2025 | Revised: 29 December 2025 | Accepted: 30 December 2025

DOI: https://doi.org/10.65114/aide.gmry6g41

Journal of Human–AI Decision-Making Systems

AI DECISIONS

The year of silent failures: AI technical debt and the crisis of verification

Yury Korolev

AI Decisions, London, UK

Correspondence

Yury Korolev, CEO, AI Decisions,
107B Cottenham Park Road,
London, SW20 0DS, UK
Email: yury@aidecisions.ai

Abstract

In recent years, generative models (LLMs) have become an "accelerator" of office work and software development: texts, reports, minutes, analytics, code, and documentation are produced and reworked faster than ever. However, this acceleration has been accompanied by the accumulation of a new class of debt — verification debt — which manifests not as immediate system failure but as organisations gradually losing the ability to verify, explain, and reproduce their decisions. We propose the conceptual framework of "knowledge rot" and "shadow metrics" for diagnosing this debt, linking the phenomenon to (i) recursive reworking of artefacts ("generation → summary → summary of summary"), (ii) the proliferation of opaque code and documents, (iii) consensus hallucinations in deliberative processes, and (iv) drift in models, data, and prompts.

The article contributes: (1) a taxonomy of mechanisms by which corporate memory degrades; (2) a formalisation of a "shadow quality circuit" by analogy with measuring the shadow economy, including a MIMIC approach to estimating the proportion of unverified AI artefacts; (3) the design of an internal "AO module" (Algorithmic Oversight module) as an independent decision-audit circuit; (4) a set of operationalisable metrics (Human Verification Rate, Data Provenance Score, Algorithmic Drift Index, Recursive Depth Index, and others) together with implementation protocols.

Keywords

Technical debt, LLM, hallucinations, corporate memory, provenance, algorithmic auditing, drift, dynamic pricing, shadow economy, MIMIC.

1. Introduction: "The hangover after the hype" and the crisis of verification

1.1 Thesis

The period 2023–2025 has been a phase of mass adoption of generative assistants: office workers have gained tools for writing, analysis, and decision preparation; developers have gained assistants for code generation. Throughput metrics and visible efficiency have improved. Yet the key question was often left unasked: **what is the quality** of this accelerated work, and **can it be verified**?

In an engineering metaphor, this resembles a skyscraper built with fast-setting concrete: growth impresses, but voids may lurk within. The crisis manifests not at the moment of pouring but later — when repair, investigation, or proof of justification is required.

1.2 Empirical background

Reports on AI adoption document rapid uptake. McKinsey's annual "State of AI" survey shows an increase in the proportion of organisations reporting the use of AI in at least one business function: from 55% in 2023 to 72% in 2024 [14]; the Stanford AI Index Report 2024 aggregates similar dynamics [13]. At the same time, these reports emphasise that only a minority of companies systematically manage risk and validate the outputs of generative models: McKinsey identifies the presence of such practices as a distinguishing characteristic of "high performers."

It is important to distinguish between **adoption success** and **knowledge-production robustness**. Rapid growth in the former without proportionate growth in the latter creates a cumulative imbalance — precisely what we call verification debt. This article focuses on the latter.

1.3 The Scientific Problem

We argue that organisations face a new form of debt:

- **Technical debt** in ML systems has been characterised as hidden future costs of maintenance and change, amplified by ML-specific factors (entanglement, hidden feedback loops, etc.) [26].
- **Verification debt** is a broader organisational phenomenon: not only code and models but also documents, decisions, minutes, and analytics become increasingly difficult to verify, reproduce, and explain.

This work contributes: (a) a theory of mechanisms, (b) measurable metrics, and (c) a control architecture.

2. Related work and normative frameworks

2.1 Technical debt in ML

The framing of "hidden technical debt" in ML systems [26] demonstrates that ML embeds data and environmental dependencies into systems and can generate massive future maintenance costs.

2.2 Model and data documentation

A line of work on documentation standards aims to make ML artefacts verifiable:

- **Model cards** [15]: standardised reporting on the behaviour and applicability boundaries of models.
- **Datasheets for datasets** [7]: documentation of provenance, composition, and limitations of datasets.
- **FactSheets / SDoC** [3]: declarations of purpose, safety, robustness, and provenance for AI services.

In this article, we extend the idea of documentation to **corporate textual artefacts** created or reworked by LLMs.

2.3 Hallucinations and LLM reliability

Surveys on LLM hallucinations [10] systematise error sources at the data, training, and inference stages and show that "plausibility" does not equal truthfulness.

A separate line of work demonstrates that increases in instructability and scale do not guarantee increased reliability on tasks where errors are difficult for humans to detect [33].

Shumailov et al. [27] demonstrated the phenomenon of **model collapse**: training on recursively generated data leads to irreversible loss of distributional tails. Although their work describes degradation at the level of model training, an analogous dynamic may manifest at the level of corporate artefacts (see §4.1).

2.4 Automation bias

Classical research in HCI [28] shows that humans tend to trust recommendations from automated systems, leading to errors of "omission" and "commission" even when incentives to verify are present.

2.5 Algorithmic auditing and risk management

The framework for internal algorithmic auditing [21] proposes an end-to-end process for the oversight of AI systems.

Normative risk-management frameworks:

- **NIST AI RMF 1.0** [16] and the **Generative AI Profile** [17].
- **ISO/IEC 42001:2023** [11] — AI management systems.
- **EU AI Act** [6], entering into force in stages from 2025: classifies AI systems by risk level and establishes obligations for transparency, data governance, and human oversight for high-risk systems. Several of its provisions — notably those on risk management, human oversight, and record-keeping — are directly relevant to verification debt (see §8.4 for a detailed mapping)

These documents provide a language of risks and controls; our aim is to operationalise them for the "verification crisis" of the LLM era.

2.6 Traceability and provenance

W3C PROV [31] establishes a standard for representing the provenance of data and artefacts; contemporary work proposes PROV-compatible models and tools for end-to-end provenance in ML pipelines [8; 24; 29].

3. Theory: Knowledge rot and verification debt

3.1 Definitions

Corporate memory — the totality of artefacts (documents, decisions, code, tickets, minutes, knowledge bases) that enable an organisation to reproduce its actions over time.

Knowledge rot — the degradation of the informativeness and verifiability of corporate memory as a consequence of recursive simplification, loss of context, and the introduction of incorrect assertions not traceable to primary sources.

Verification debt — the accumulated gap between (i) the volume of decisions and artefacts produced by the "human + LLM" system and (ii) the organisation's actual capacity to verify their correctness, justification, and provenance.

3.2 Why the "Debt" compounds

Whereas in classical technical debt the proportion of "incomprehensible code" grows roughly linearly with the volume of changes, with LLMs **positive feedback loops** emerge:

1. Accelerated generation → more artefacts → less time for verification.
2. The appearance of "second-order" artefacts (summaries, roadmaps, resolutions) that begin to live independently of primary sources.
3. Increased trust in presentation (fluent text) raises the probability of accepting an error — **fluency bias**.
4. Invisible drift in models and data (§4.4) adds non-stationarity to an already opaque process.

3.3 Hypotheses

We formulate testable hypotheses (in terms of corporate data):

- **H1 (Compression drift):** as the proportion of LLM-generated summaries increases, the divergence between summary content and primary sources grows, measured by errors of omission of key conditions, implicit assumptions, and referential integrity.
- **H2 (Understanding gap):** as the proportion of LLM-generated code increases, the time to localise defects and the proportion of "re-prompting" (rather than systematic debugging) grow.
- **H3 (Consensus hallucination):** automated meeting minutes systematically understate the intensity of conflict and the degree of decision uncertainty (compared with audio/transcripts).
- **H4 (Audit asymmetry):** organisations with an independent oversight circuit (AO module) exhibit a lower rate of verification-debt accumulation at comparable generation rates.

4. Mechanisms of degradation

4.1 "Chinese whispers": Recursive compression

Scenario: an employee uses an LLM to write a report; another uses an LLM to summarise the report; a third produces an "executive summary"; then a presentation and a decision protocol emerge.

Key mechanism: each step is, in essence, lossy compression with the addition of probabilistic errors. Even if each individual transformation introduces only a small error, the composition of transformations leads to systematic loss of:

- context (why a decision was taken),
- conditions of applicability (under what constraints the decision is valid),
- alternatives and conflicts,
- references to primary data.

Operationalisation: measure "reference integrity" — the proportion of assertions in a summary that can be mapped to specific spans in the primary source (transcript span alignment) and/or data. Additionally, measure Recursive Depth Index (see M7, §7).

Link to LLM risks: surveys on hallucinations [10] show that models can produce plausible but incorrect assertions; increases in instructability do not guarantee a reduction in such errors [33].

Link to model collapse: Shumailov et al. [27] showed that training models on recursively generated data leads to "model collapse" — the irreversible loss of distributional tails. Although their work describes degradation at the level of model training, an analogous dynamic manifests at the level of corporate artefacts: each "summary of a summary" trims the tails — rare but critically important conditions, warnings, and nuances. We term this **organisational knowledge-distribution collapse**.

4.2 Code nobody understands: The operatorisation of development

Scenario: junior engineers use Copilot/LLM to write working code quickly. It passes basic tests but contains hidden vulnerabilities, sub-optimalities, and architectural "seams." During an incident, the team cannot explain why the code is structured as it is and begins to "prompt" a solution instead of performing systematic debugging.

Empirical grounding: research on the security of AI coding assistants [19] shows that a substantial proportion of generated fragments can be vulnerable in high-risk scenarios.

Mechanics of the debt:

- growth of "local patches" instead of architectural solutions;
- deterioration of developers' cognitive model of the system (understanding is replaced by operational skill);
- increased cost of change (change amplification).

Instrumental diagnostics:

- proportion of PRs in which key blocks are "explained" by reference to the LLM;
- proportion of critical modules lacking human-authored comments on purpose and invariants;
- security metrics (CWE) on a sample of LLM suggestions.

4.3 The illusion of agreement: consensus hallucinations

Scenario: AI agents produce meeting minutes and formulate "action items." For readability, they smooth over conflict and ambiguity.

Risk: management sees a tidy picture ("everything agreed"), whereas the actual discussions contained warnings, disagreements, or conditions.

Path to measurement: comparison of the protocol with the transcript and construction of indicators:

- *Conflict Signal Loss:* the proportion of utterances expressing disagreement or flagging risks that are absent from the protocol.
- *Uncertainty Flattening:* the reduction of uncertainty markers (may/if/subject to) in the final text.

In the field of meeting summarisation, datasets and methods exist that enable such checks (e.g. QMSum [32]), as well as work on post-hoc correction of summaries using multi-LLM approaches [12].

4.4 The invisible shift: drift in models, data, and prompts

Scenario: an organisation launched an AI decision-support system (or report-generation system) in Q1. By Q3 the provider had updated the base model, the team had changed the system prompt four times, and the distribution of input data had shifted owing to seasonality or changes in business processes. None of these changes was documented as a "release"; the cumulative effect on outputs is unknown.

Mechanisms of drift:

- **Data drift:** a change in the distribution of input data, which may push the model beyond its training domain.
- **Concept drift:** a change in the target variable with features unchanged (e.g. what counts as "normal" customer behaviour changes).
- **Model/API drift:** the provider updates the base model without notification or with a minimal changelog.
- **Prompt drift:** cumulative edits to prompts without version control or regression testing.

Compounding: each type of drift in isolation may be minor, but their product creates a "space of unpredictability" in which system behaviour diverges from expectations without an overt failure signal.

Operationalisation: Algorithmic Drift Index (ADI; see §7, M3) and mandatory "checkpoints" upon any change to model, prompt, or data (see §8).

Literature: surveys on drift detection [9] describe detection methods, including for unsupervised settings. For LLM systems, an additional complication is that "drift" may be indistinguishable from "improvement" without an external benchmark.

5. Measurement: from the "Shadow economy" to "Shadow quality"

5.1 Analogy and motivation

The shadow economy is the portion of economic activity hidden from official measurement. Latent models are applied to it (e.g. MIMIC [1; 4; 25]), in which the "true size" is not directly observable but manifests through indicators and causes.

We propose an analogue: the **shadow quality circuit** — the proportion of corporate artefacts and decisions for which sufficient evidence of correctness, provenance, and verification is absent. These are not necessarily "bad" artefacts; they are artefacts of **unknown quality** that increase the risk of systemic failures.

5.2 Defining the latent variable

Let Q* be the latent index of "verification opacity" per unit of time (e.g. one week) at the level of a team or business unit.

We wish to estimate Q* from observable indicators (MIMIC):

- **Indicators:** symptoms of shadow quality.
- **Causes:** structural factors that increase Q*.

5.3 Indicators (examples)

1. **Low traceability ratio:** the proportion of assertions in key documents not traceable to a primary source (no citation, no span-alignment, no data).
2. **Unreviewed AI output share:** the proportion of LLM artefacts that have not undergone human review (see HVR).
3. **Regression-to-fluency:** an increase in linguistic fluency accompanied by a decline in factual consistency (e.g. as measured by automatic factuality metrics).
4. **Incident explainability gap:** the proportion of incidents in which root-cause analysis (RCA) cannot reconstruct the decision chain.
5. **Code comprehension lag:** the time to fix a defect in a module with a high proportion of LLM-generated code.

5.4 Causes (examples)

1. **AI throughput pressure:** an increase in the number of artefacts per person.
2. **Low reviewer capacity:** a low ratio of reviewer-hours to generated-hours.
3. **Weak provenance infrastructure:** the absence of PROV-compatible tracing.
4. **Incentive misalignment:** KPIs for speed/volume without KPIs for verifiability.
5. **Model/Prompt volatility:** frequent changes to models/prompts without drift control.

5.5 Mapping MIMIC variables to operationalised metrics

Table 1. Correspondence of MIMIC indicators/causes to metrics in §7

Role in MIMIC	Variable	Metric (§7)
Indicator	Low Traceability Ratio	M4 (RIR)
Indicator	Unreviewed AI Output Share	M1 (HVR, inverse)
Indicator	Regression-to-Fluency	— (requires factuality scorer)
Indicator	Incident Explainability Gap	M5 (DRR, inverse)
Indicator	Code Comprehension Lag	— (measured via MTTR)
Cause	AI Throughput Pressure	Artefacts/person/week
Cause	Low Reviewer Capacity	1 − HVR
Cause	Weak Provenance Infrastructure	M2 (DPS, inverse)
Cause	Incentive Misalignment	Share of KPIs for speed vs. verifiability
Cause	Model/Prompt Volatility	M3 (ADI)

This explicit correspondence enables the construction of a MIMIC model directly from operationalised metrics.

5.6 MIMIC specification

The standard MIMIC model is specified by two equations:

Structural equation (causes → latent):

$$Q^* = \gamma_1 C_1 + \gamma_2 C_2 + \ldots + \gamma_k C_k + \zeta$$

Measurement equations (latent → indicators):

$$I_j = \lambda_j Q^* + \varepsilon_j, \text{ for } j = 1 \ldots m$$

where C_i are causes, I_j are indicators, and ζ and ε_j are error terms.

5.7 Identification and calibration

As in the measurement of the shadow economy, the absolute scale of Q* requires normalisation. Possible approaches:

- normalisation to a "reference" period prior to LLM adoption;
- normalisation across units with differing intensities of LLM use;
- calibration against an audit sample (manual quality assessment) as an "anchor."

5.8 Why this matters

Shadow indicators allow organisations not to wait for a catastrophe (incident) but to track **precursors** of corporate-memory degradation.

6. Case study: "Anatomy of a failure" in Dynamic pricing

6.1 Setting

Consider a constructed but plausible case of a large retailer that has deployed AI for dynamic pricing.

Objective: optimisation of short-term profit (e.g. margin) subject to constraints.

Artefacts:

- demand and elasticity model,
- customer-segmentation rules,
- reports on profit growth,
- decision protocols on experiment expansion,
- risk and compliance documents.

6.2 The hidden error

The model optimises short-term profit but does not account for long-term effects: loyalty, LTV, fairness perception, and regulatory risk.

An aggressive policy leads to loyal customers (low churn probability) being offered higher prices.

The literature on personalised/discriminatory pricing [18; 20; 23] shows that perceptions of unfairness can reduce utility and undermine platform sustainability; it also shows that different bases for personalisation are perceived as more or less fair.

6.3 The dynamics of "Success"

- Q1–Q3: profit grows; KPIs are met.
- Q4: mass churn begins, NPS falls, support contacts surge.
- Signals were present in the data (returns, complaints, "price shock"), but protocols and reports smoothed over the risks.

6.4 Episodes of external validity

The constructed case draws empirical support from several documented episodes:

1. **Instacart (late 2025):** according to Reuters and AP News [2; 22], Instacart discontinued its programme of algorithmic price testing following criticism: users could see different prices for the same item at the same store.
2. **Amazon (2000):** one of the earliest documented cases of personalised pricing — Amazon tested different prices for DVDs for different users, leading to a public scandal and an apology from the company [30].

3. **Uber surge pricing:** years of criticism of algorithmic pricing, including cases of extreme surge pricing during emergencies, illustrate the gap between metric optimisation and public perceptions of fairness [5].

All three episodes illustrate the **gap between metric optimisation and trust/transparency**, even when personal data were not formally used.

6.5 The moral of the case

The algorithm could "hit KPIs" yet harm the business. Organisations therefore need a circuit external to the model:

- monitoring of fairness/trust,
- provenance control for decisions,
- independent sample-based verification.

We call this the **AO module**.

7. Shadow quality metrics: Specification, Calculation, and Risk thresholds

7.1 Why traditional KPIs "lie"

ROI and "efficiency" capture speed and short-term effect but are poor at detecting:

- loss of context,
- degradation of reproducibility,
- growth of risk,
- incentive distortion.

7.2 Core metric set

Below is a minimal set of metrics for "knowledge-system health." Threshold values must be calibrated to context; the thresholds given here are **heuristics for triage**, not normative standards.

(M1) Human Verification Rate (HVR)

Definition: the proportion of AI artefacts for which a recorded human review event exists at the level of content (not format alone).

Measurement:

HVR = (number of AI outputs with review) / (number of AI outputs produced)

Risk: a low HVR means the organisation produces more "unknown quality" than it is able to verify.

(M2) Data Provenance Score (DPS)

Definition: the degree to which assertions and figures can be traced to data and sources.

Components:

- presence of a source reference,
- presence of a data-version identifier,
- presence of a transformation chain (PROV),
- ability to reproduce the calculation.

Example scale (0–100):

- 0–20: "source unknown"
- 20–50: reference exists but no version/calculation
- 50–80: partial reproducibility
- 80–100: end-to-end reproducibility

(M3) Algorithmic Drift Index (ADI)

Definition: the degree of change in model behaviour relative to the baseline at deployment.

Components:

- data drift,
- concept drift,
- prompt and post-processing drift,
- model/weight updates.

Methods from the drift-detection literature [9] and organisational "checkpoints" should be used.

(M4) Reference Integrity Ratio (RIR)

Definition: the proportion of assertions in a document that can be linked to a primary source.

Formally:

RIR = (number of claims with evidence link) / (number of claims total)

An "evidence link" may take the form of:

- a citation with coordinates in the document,
- a span in a transcript,
- a query to a database/dashboard,
- a reference to a ticket/PR.

(M5) Decision Reproducibility Rate (DRR)

Definition: the proportion of decisions for which the following can be reconstructed: (i) alternatives, (ii) arguments, (iii) data, (iv) persons, (v) conditions.

(M6) Security Risk in AI Code (SRAC)

Definition: the proportion of AI code suggestions containing high-risk CWE patterns under static analysis and review.

(M7) Recursive Depth Index (RDI)

Definition: the mean number of generative steps between the current artefact and the nearest primary source (raw data, original document, audio recording).

Measurement:

RDI = (1 / |D|) × Σ depth(d), for all d in D

where depth(d) is the number of intermediate AI transformations from the primary source to artefact d.

Example: primary report (depth = 0) → AI summary (depth = 1) → executive summary of the AI summary (depth = 2) → decision protocol based on the executive summary (depth = 3).

Risk: for critical decisions, when RDI > 2 the probability of losing material information increases non-linearly (hypothesis H1).

7.3 Risk matrix by metric (example)

Table 2. Risk matrix by metric (example)

Level	HVR	DPS	ADI	RIR	DRR	RDI (critical)
Normal	> 30%	> 70	low / controlled	> 0.7	> 0.6	≤ 1
Warning	10–30%	40–70	medium	0.4–0.7	0.3–0.6	2
Risk	< 10%	< 40	high / unknown	< 0.4	< 0.3	≥ 3

Note: thresholds depend on domain criticality. In medicine and finance, substantially higher levels are required.

8. Control architecture: The "AO module" and demonstrable traceability

8.1 The principle of the independent circuit

The AO module (Algorithmic oversight) is an independent function or service that:

1. performs sample-based verification of AI artefacts and decisions;
2. measures metrics (§7);
3. ensures provenance (§8.2);
4. initiates investigations and corrective actions.

It is the organisational analogue of "internal audit" and "quality control," adapted for probabilistic systems.

8.2 Provenance as infrastructure

Reproducibility requires recording chains:

data → transformations → features → model → decision → document → action.

The standard language used is W3C PROV [31] (PROV-DM/PROV-O). Contemporary tools offer PROV-compatible models for ML pipelines and automated provenance extraction (e.g. MLflow2PROV [8; 24]).

Minimum provenance specification for LLM artefacts:

- model identifier/version,
- configuration (system prompt, policies, tools),
- input data (with references and versions),
- output artefact (hash),
- human reviewer (who reviewed, what was reviewed),
- context of use (decision/ticket),
- recursive depth (RDI).

8.3 Sample-based audit and "Control groups"

To avoid being overwhelmed by volume, a statistical approach is required:

- stratified sampling by criticality;
- "red-team" checks for hallucinations/leaks;
- control groups without LLM use (or with limited use) for causal-effect estimation.

Minimum sample-based verification protocol (AO module):

1. Define strata by criticality (finance / security / compliance / operations).
2. Weekly, select n artefacts from each stratum.
3. For each artefact:
    - check RIR (presence of evidence links);
    - verify 5–10 factual assertions;
    - reconcile 1–2 key figures to source data;
    - assess RDI; where depth ≥ 3, mandatory verification to the primary source;
    - record findings and corrective actions.

8.4 Compatibility with normative frameworks

- **NIST AI RMF** establishes the functions GOVERN / MAP / MEASURE / MANAGE; the metrics in §7 correspond to MEASURE and MANAGE; provenance and audit correspond to GOVERN.
- **ISO/IEC 42001:2023** [11] establishes an AI management system; the AO module implements controls (Annex A, A.6 "Operation of AI system," A.7 "Performance evaluation") and the continuous-improvement mechanism (§10 of the standard).
- **EU AI Act** (Regulation (EU) 2024/1689) [6]: for high-risk AI systems (Annex III) the AO module can serve as an instrument for fulfilling obligations under:
    - Art. 9 (risk management system),
    - Art. 11 (technical documentation),

- Art. 12 (record-keeping),
- Art. 14 (human oversight),
- Art. 15 (accuracy, robustness, cybersecurity).

For general-purpose AI models (GPAI, Chapter V), Article 53 requires technical documentation and transparency policies — the DPS and RIR metrics are directly relevant.

9. Roadmap: What to do in 2026

9.1 Inventory (0–4 Weeks)

1. Catalogue of artefacts (where LLMs are used: documents, code, minutes, analytics).
2. Identification of "critical circuits" (pricing, compliance, security, finance).
3. Rapid calculation of baseline metrics (HVR, DPS, RIR, RDI) on a pilot domain.

9.2 Instrumentation (1–3 Months)

1. Logging of LLM sessions (with due regard for privacy): metadata, versions, hashes.
2. Implementation of simple evidence-link mechanisms (document templates, mandatory reference fields).
3. Automated drift monitoring (minimum: data drift + prompt/model changes).

9.3 Establishing the AO module (3–6 Months)

1. Roles and authority (independence from development teams).
2. Regulations for sample-based audit.
3. Metrics as quality KPIs (not speed KPIs).

9.4 Cleaning corporate memory (6–12 Months)

1. "Decompression" of key knowledge: restoration of primary sources.
2. Mandatory "artefact cards": document and code cards by analogy with Model Cards.

Table 3. "LLM artifact card" template for corporate documents

Field	Description	Example
1. Purpose	What the document was created for; decisions it influences	"Analytical brief for the pricing committee"
2. Data sources and primary sources	List of primary sources + data versions + extraction dates	pricing_db v3.2, extraction 2025-01-15; CRM export #4412
3. LLM configuration	Model/version; system prompt (brief); tools (RAG, code interpreter, API)	Claude 3.5 Sonnet (2025-10-22); prompt v2.1; RAG → pricing_kb

Field	Description	Example
4. Recursive depth (RDI)	Number of intermediate AI transformations from primary source	RDI = 1 (direct summarisation of report)
5. Verification	Who reviewed; what was reviewed (facts / figures / logic / completeness); reconciliations	A. Ivanov, 2025-01-20: 12 key figures reconciled against pricing_db; 2 discrepancies corrected
6. Limitations	Where the document must NOT be used; known simplifications	Not to be used for regulatory filings without additional legal review
7. Validity period	Date after which the document requires revision	2025-04-15 (quarterly cycle)

3. Archival of "unverified" materials as drafts, not as "truth."

10. Limitations and research programme

10.1 Limitations

- Metrics require access to corporate data and may be sensitive.
- Provenance infrastructure is complex and has implementation costs.
- The danger of **Goodhart's Law**: metrics may become targets and lose their diagnostic value (e.g. formal "ticking of the box" in HVR without genuine review).
- **The case study (§6) is constructed**, not based on full access to an organisation's data. Empirical validation of hypotheses H1–H4 requires partnerships with organisations or access to logs of real AI pipelines.
- **Threshold values (§7.3)** are based on expert heuristics, not empirical calibration. They should be treated as starting points for adaptation, not as normative standards.
- **Measurement problem:** determining which artefacts are "AI-generated" becomes increasingly difficult in a regime of co-authorship (human + AI); boundaries are blurred.

10.2 Research programme

1. Causal estimation of LLM impact on decision quality (DiD, event study, RCT at the team level).
2. Experimental models of degradation under recursive summarisation — quantitative estimation of losses over N steps (extending Shumailov et al. [27] to the textual domain).
3. Design of organisational incentives to increase HVR/DPS (behavioural experiments).
4. Development of "LLM Artifact Card" standards for documents and minutes.
5. **Automated verification:** investigation of the capabilities of multi-agent systems for cross-checking AI artefacts (LLM-as-judge approaches) and the limits of their reliability.
6. **Longitudinal studies:** tracking of metrics M1–M7 in organisations over a 12–24-month horizon for empirical testing of the compounding hypothesis (§3.2).

11. Conclusion

Organisations have already deployed LLMs into the production of texts and code. The next stage is the transition from the question "how to deploy more AI?" to the question **"how to verify what the AI has already done?"**

The problem does not require futurology: it is a tractable engineering-and-organisational challenge. What is needed is metrics, provenance, an independent audit circuit, and a discipline of "demonstrable reproducibility."

Conflict of interest: The authors declare no conflict of interest

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work, the author(s) used AI tool, namely Claude, in order to correct Grammatical mistakes and edit the language professionally. After using this tool/service, the author(s) reviewed and edited the content as needed.

ORCID

Yury Korolev https://orcid.org/0000-0001-8316-0058

Decision Impact Summary

This article supports decisions by technology leaders, risk officers, and AI governance teams about how to manage the growing volume of AI-generated corporate artefacts—documents, code, meeting minutes, and analytics. The proposed framework argues that rapid LLM adoption creates a compounding "verification debt": organisations produce more than they can verify, and traditional KPIs for speed and efficiency fail to detect the resulting erosion of decision quality, traceability, and reproducibility. Practitioners should therefore treat verification capacity as a first-class operational constraint: inventory where LLMs touch critical decisions, instrument those workflows with the proposed shadow-quality metrics (Human Verification Rate, Data Provenance Score, Algorithmic Drift Index, Reference Integrity Ratio, Recursive Depth Index), and establish an independent Algorithmic Oversight module that performs sample-based audits stratified by criticality. Human oversight remains essential—every high-stakes artefact requires a recorded content-level review, and provenance chains from data to decision should be logged using standard formats (W3C PROV). The principal risks are recursive information loss ("summary of a summary"), consensus hallucinations in automated minutes, undetected model and prompt drift, and fluency bias that masks factual errors. These are best addressed through mandatory evidence-linking in document templates, version control for prompts and models, and KPIs that reward verifiability alongside throughput. The paper provides an actionable 12-month roadmap, a set of operationalisable metrics with triage thresholds, and an "LLM Artifact Card" template, enabling organisations to begin measuring and reducing verification debt immediately rather than waiting for a costly failure.

References

[1] Almenar, V., Sánchez, J. L., Sapena, J. Measuring the shadow economy and its drivers. *Economic Research–Ekonomska Istraživanja* (2020). DOI: 10.1080/1331677X.2019.1706601. URL: https://www.tandfonline.com/doi/full/10.1080/1331677X.2019.1706601
[2] AP News. Instacart ends a program where users could see different prices for the same item at the same store. (2025-12-22). URL: https://apnews.com/article/c9a0a52e959ce46d2152aa664308d228 (accessed 24 December 2025).
[3] Arnold, M., Bellamy, R. K. E., Hind, M., et al. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. arXiv:1808.07261. URL: https://arxiv.org/abs/1808.07261
[4] Buehn, A., Schneider, F. Shadow economies around the world: novel insights, accepted knowledge, and new estimates. *International Tax and Public Finance* (2012). DOI: 10.1007/s10797-011-9187-7. URL: https://link.springer.com/article/10.1007/s10797-011-9187-7
[5] Chen, M. K., Mislove, A., Wilson, C. Peeking Beneath the Hood of Uber. *Proc. ACM IMC 2015*. DOI: 10.1145/2815675.2815681. URL: https://doi.org/10.1145/2815675.2815681
[6] European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). OJ L, 2024/1689 (2024-07-12). URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj
[7] Gebru, T., Morgenstern, J., Vecchione, B., et al. Datasheets for Datasets. *Communications of the ACM* (2021). DOI: 10.1145/3458723. URL: https://dl.acm.org/doi/10.1145/3458723
[8] Grafberger, S. Provenance Tracking for End-to-End Machine Learning Pipelines. *WWW '23 Companion*. DOI: 10.1145/3543873.3587557. URL: https://dl.acm.org/doi/10.1145/3543873.3587557
[9] Hinder, F., et al. One or two things we know about concept drift — a survey on drift detection (unsupervised setting). *Frontiers in Artificial Intelligence* (2024). DOI: 10.3389/frai.2024.1330257. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11220237/
[10] Huang, L., Yu, W., et al. A Survey on Hallucination in Large Language Models. *ACM Computing Surveys* (2025). DOI: 10.1145/3703155. URL: https://dl.acm.org/doi/10.1145/3703155
[11] ISO. ISO/IEC 42001:2023 — AI management systems. URL: https://www.iso.org/standard/81230.html
[12] Kirstein, F., et al. What's Wrong? Refining Meeting Summaries with LLM Feedback. arXiv:2407.11919. URL: https://arxiv.org/abs/2407.11919
[13] Maslej, N., Fattorini, L., Perrault, R., et al. The AI Index 2024 Annual Report. Stanford HAI (2024). URL: https://aiindex.stanford.edu/report/
[14] McKinsey & Company. The state of AI in early 2024: Gen AI adoption spikes and starts to generate value. (2024-05). URL: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai (accessed 15 December 2025).
[15] Mitchell, M., Wu, S., Zaldivar, A., et al. Model Cards for Model Reporting. *FAT* 2019*. DOI: 10.1145/3287560.3287596. URL: https://dl.acm.org/doi/10.1145/3287560.3287596
[16] NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1 (2023). URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
[17] NIST. AI RMF: Generative AI Profile. NIST AI 600-1 (2024). DOI: 10.6028/NIST.AI.600-1. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[18] OECD. Personalised Pricing in the Digital Era. Background note, DAF/COMP(2018)13 (2018). URL: https://www.oecd.org/en/publications/personalised-pricing-in-the-digital-era_fd24f09b-en.html
[19] Pearce, H., Ahmad, B., Tan, B., Dolan-Gavitt, B., Karri, R. Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions. *IEEE S&P 2022*. DOI: 10.1109/SP46214.2022.9833571. URL: https://arxiv.org/abs/2108.09293
[20] Priester, A., Robbert, T., Roth, S. A special price just for you: effects of personalized dynamic pricing on consumer fairness perceptions. *Journal of Revenue and Pricing Management* (2020). DOI: 10.1057/s41272-019-00224-3. URL: https://link.springer.com/article/10.1057/s41272-019-00224-3
[21] Raji, I. D., Smart, A., White, R. N., et al. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *FAccT 2020*. DOI: 10.1145/3351095.3372873. URL: https://dl.acm.org/doi/10.1145/3351095.3372873
[22] Reuters. Instacart ends AI-driven price experiments after criticism. (2025-12-22). URL: https://www.reuters.com/business/instacart-ends-ai-driven-price-experiments-after-criticism-2025-12-22/ (accessed 22 December 2025).
[23] Richards, T. J., Liaukonyte, J., Streletskaya, N. A. Personalized pricing and price fairness. *International Journal of Industrial Organization* (2016). DOI: 10.1016/j.ijindorg.2015.11.004. URL: https://www.sciencedirect.com/science/article/pii/S0167718715001216
[24] Schlegel, M., et al. Capturing end-to-end provenance for machine learning pipelines. *Information Systems* (2025). URL: https://www.sciencedirect.com/science/article/pii/S0306437924001534
[25] Schneider, F., Enste, D. Shadow Economies: Size, Causes, and Consequences. *Journal of Economic Literature* (2000). DOI: 10.1257/jel.38.1.77. URL: https://www.aeaweb.org/articles?id=10.1257/jel.38.1.77
[26] Sculley, D., Holt, G., Golovin, D., et al. Hidden Technical Debt in Machine Learning Systems. *NeurIPS 2015*. URL: https://proceedings.neurips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html
[27] Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., Anderson, R. AI models collapse when trained on recursively generated data. *Nature* 631, 755–759 (2024). DOI: 10.1038/s41586-024-07566-y. URL: https://www.nature.com/articles/s41586-024-07566-y
[28] Skitka, L. J., Mosier, K., Burdick, M. Does automation bias decision-making? *International Journal of Human-Computer Studies* (1999). DOI: 10.1006/ijhc.1999.0252. URL: https://doi.org/10.1006/ijhc.1999.0252
[29] Souza, R., Valduriez, P., Mattoso, M. Provenance Data in the Machine Learning Lifecycle in Computational Science and Engineering. *WORKS 2019*. DOI: 10.1109/WORKS49585.2019.00006. URL: https://doi.org/10.1109/WORKS49585.2019.00006
[30] Streitfeld, D. "On the Web, Price Tags Blur." *The Washington Post*, 27 September 2000. (Accessed via newspaper archives; original URL may require subscription.)
[31] W3C. PROV-Overview / PROV-DM / PROV-O (Recommendations). URL: https://www.w3.org/TR/prov-overview/ and https://www.w3.org/TR/prov-o/
[32] Zhong, M., Yin, D., Yu, T., et al. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. *Proceedings of NAACL-HLT 2021*, pp. 5905–5921. arXiv: https://arxiv.org/abs/2104.05938
[33] Zhou, L., Schellaert, W., Martínez-Plumed, F., et al. Larger and more instructable language models become less reliable. *Nature* (2024). DOI: 10.1038/s41586-024-07930-y. URL: https://www.nature.com/articles/s41586-024-07930-y