



Baština Akademije nauka i umjetnosti Bosne i Hercegovine

## **SMART MANUFACTURING ON THE HORIZON: A Conceptual Shift from Industry 4.0 to Industry 5.0: Industrija**

**Karabegović, Isak**

**2026**

Academy of Sciences and Arts of Bosnia and Herzegovina

<https://bastina.anubih.ba/handle/123456789/875>

Preuzeto s Baštine Akademije nauka i umjetnosti Bosne i Hercegovine

<https://bastina.anubih.ba/>

## Semantic Data Pollution: A New Challenge For Cybersecurity and Standardization of AI Models in Smart Factories

Samir Lemeš<sup>\*1</sup>

**Abstract:** *The transition from Industry 4.0 to Industry 5.0 introduces a fundamental shift from data-intensive automation toward human-centred, resilient, and trustworthy intelligent manufacturing systems. While industrial artificial intelligence increasingly relies on synthetic data generated by digital twins, autonomous reporting systems, and interconnected AI agents, recursive retraining on machine-generated outputs creates a previously underexplored vulnerability. This paper introduces the concept of Semantic Data Pollution, defined as the progressive degradation of semantic integrity when AI systems increasingly learn from synthetic representations of reality instead of authentic physical observations. Drawing upon recent research on model collapse in generative AI, the paper demonstrates how recursive learning gradually eliminates statistically rare but operationally critical anomalies, reducing the ability of industrial AI systems to detect machine failures and cyberattacks. The problem is analysed within Edge–Cloud smart factory architectures where local semantic corruption may propagate system-wide through centralized retraining pipelines. To address this challenge, a multi-layer mitigation framework is proposed based on data provenance tracking, synthetic data watermarking, statistical monitoring of distribution drift, and Human-in-the-Loop validation. The paper argues that future industrial cybersecurity standards must evolve beyond conventional confidentiality and integrity protection toward preserving the semantic integrity of AI-driven industrial decision systems.*

**Keywords:** *artificial intelligence, semantic pollution, recursive learning, model collapse*

### 1. Introduction

The rapid development of industrial digitalization over the last decade has fundamentally transformed manufacturing systems, creating the technological foundations of what has been widely recognized as Industry 4.0. The central paradigm of Industry 4.0 focused primarily on extensive deployment of cyber-physical systems, Industrial Internet of Things (IIoT), autonomous communication networks, and large-scale data acquisition infrastructures intended to maximize machine connectivity and generate vast quantities of operational data. In this paradigm, data quantity was treated as a primary strategic

---

<sup>\*1</sup>University of Zenica, Polytechnic Faculty, Zenica, Bosnia and Herzegovina  
E-mail: [samir.lemes@unze.ba](mailto:samir.lemes@unze.ba)

asset, giving rise to the Big Data approach in industrial automation, where the assumption prevailed that more data inherently leads to better optimization, prediction accuracy, and decision support.

The emerging transition toward Industry 5.0 introduces a significant conceptual shift. Rather than prioritizing automation efficiency and unrestricted data accumulation, Industry 5.0 emphasizes human-centric production systems, resilience, sustainability, and collaborative interaction between artificial intelligence systems and human operators. Within this framework, data quality, contextual understanding, and semantic reliability become substantially more important than mere data volume. This transition creates a previously underexplored vulnerability in industrial AI systems: the progressive degradation of semantic integrity in machine learning environments increasingly dependent on recursively generated synthetic data.

This paper introduces the concept of semantic data pollution, defined as the gradual contamination of industrial information ecosystems through recursive circulation of machine-generated semantic artifacts that progressively diverge from real-world operational conditions while preserving superficial structural validity. Unlike conventional cybersecurity threats, which primarily target confidentiality, integrity, and availability of information systems [1], semantic data pollution represents a degradation of meaning itself. The threat emerges when artificial intelligence systems deployed in industrial environments begin learning from outputs generated by other AI systems rather than from authentic observations of physical processes.

A practical example can be observed in predictive maintenance systems deployed in smart factories. Modern predictive maintenance architectures continuously analyze sensor data streams to identify anomalous machine behavior and predict equipment failure before breakdown occurs. However, as AI systems increasingly automate reporting, anomaly classification, maintenance recommendations, and incident summarization, newly generated machine learning models may begin training on datasets partially composed of outputs created by earlier AI systems. Over successive learning cycles, these recursive feedback loops introduce systematic distortion into the training distribution.

The principal danger lies in what recent machine learning literature identifies as heavy tails [2]. Rare events, abnormal states, unexpected machine failures, and edge-case operational anomalies occupy statistically sparse regions of the original data distribution. Recursive training on synthetic outputs progressively removes these low-frequency observations because generative systems tend to reproduce dominant statistical patterns while underrepresenting exceptional conditions. Eventually, industrial AI systems may lose the capacity to recognize precisely those rare anomalies they were originally designed to detect.

This phenomenon has recently attracted considerable attention in the broader artificial intelligence research community under the concept of Model Collapse.

This concept was formally described by Shumailov et al. in [3]. They demonstrated that generative models recursively trained on synthetic data progressively lose information located in the tails of the original probability distribution, eventually converging toward increasingly distorted representations of reality. Their work showed that early-stage collapse first affects rare events, while advanced collapse leads to severe degradation of model diversity and representational fidelity.

Although this phenomenon has primarily been studied in the context of large language models and generative AI systems, its implications for industrial environments may be considerably more severe. Unlike consumer-facing AI applications, industrial AI systems directly interact with physical infrastructure, critical production processes, and safety-sensitive operational environments. A predictive maintenance model that gradually loses sensitivity to rare but dangerous machine failure signatures introduces a new category of cyber-physical vulnerability that cannot be detected through conventional cybersecurity frameworks.

Statistical analysis by Seddik et al. [4] further demonstrated the inevitability of model degradation under recursive synthetic-data training conditions. Their mathematical framework proved that systems trained exclusively on synthetic data inevitably experience progressive information loss, while stable learning can only be maintained when sufficient proportions of authentic human-generated or real-world data remain continuously integrated into training pipelines. This result has profound implications for industrial AI architectures increasingly dependent on automated data labeling and synthetic data augmentation processes.

Parallel developments within industrial AI research have increasingly emphasized semantic modeling as a critical foundation for trustworthy manufacturing intelligence systems. Martínez-Arellano and Ratchev [5] proposed semantic models as an architectural layer intended to improve industrial AI reusability, interoperability, and contextual machine understanding in manufacturing environments. Their work highlights the growing dependence of industrial AI systems on semantically structured knowledge representations, suggesting that semantic degradation may directly compromise operational reliability of future autonomous factories.

Complementary concerns regarding industrial data quality were systematically analyzed in the recent meta-review by researchers investigating data-related failures in industrial AI implementation [6]. Examining seventy-two distinct categories of industrial AI data issues, the study demonstrated that data reliability problems increasingly dominate operational AI failures, particularly during data integration, preprocessing, interoperability, and security phases. The findings indicate that future industrial AI failures may arise less from algorithmic limitations and more from degradation of data quality throughout complex industrial data pipelines.

Further expansion of the model collapse problem into multi-agent AI environments was investigated by Hu et al. [7], who examined recursive synthetic-data interactions across multimodal generative systems. Their findings demonstrate that model collapse becomes significantly more complex when multiple autonomous AI systems interact and exchange synthetic outputs, closely resembling future smart factory architectures where distributed AI agents independently monitor production lines, coordinate maintenance schedules, and exchange operational intelligence. Their work suggests that autonomous industrial ecosystems may be particularly vulnerable to recursive semantic degradation unless corrective mechanisms are implemented.

Taken together, these recent studies reveal the emergence of a new class of industrial vulnerability extending beyond traditional cybersecurity definitions. Existing industrial cybersecurity frameworks, including standards derived from International Electrotechnical Commission (IEC) and International Organization for Standardization (ISO) architectures [8], focus primarily on preventing unauthorized access, protecting communication channels, ensuring data confidentiality, and maintaining system availability. However, none adequately address the problem of semantic integrity degradation caused by recursive AI-generated information contamination.

Within Industry 5.0, the role of the human operator therefore becomes strategically critical. Human experts function not merely as supervisors but as semantic validators capable of distinguishing authentic operational anomalies from recursively generated synthetic artifacts. This elevates the concept of Human-in-the-Loop from an operational convenience to a fundamental resilience mechanism against semantic degradation of industrial intelligence systems.

This paper argues that future industrial cybersecurity standards must expand beyond traditional protection models and incorporate semantic data cybersecurity, defined as the protection of meaning, contextual fidelity, and epistemic integrity within AI-driven industrial decision systems. Semantic data pollution represents an emerging systemic threat to autonomous factories, and failure to address it may lead to gradual but potentially catastrophic degradation of industrial AI reliability despite the apparent preservation of conventional cybersecurity protections.

## 2. Vulnerability Vector

The rapid advancement of artificial intelligence has introduced an unprecedented paradox in the evolution of digital systems. As AI systems become increasingly capable of generating high-quality synthetic content, they also begin producing the very data on which future generations of AI systems are trained. This recursive relationship between generation and learning creates a previously underestimated phenomenon that may be described as Semantic Data Pollution: the gradual degradation of information quality when artificial intelligence systems

increasingly learn from synthetic representations of reality rather than reality itself. Recursion is a direct threat to the physical integrity of production, because an AI learns from its own hallucinations (Fig. 1).

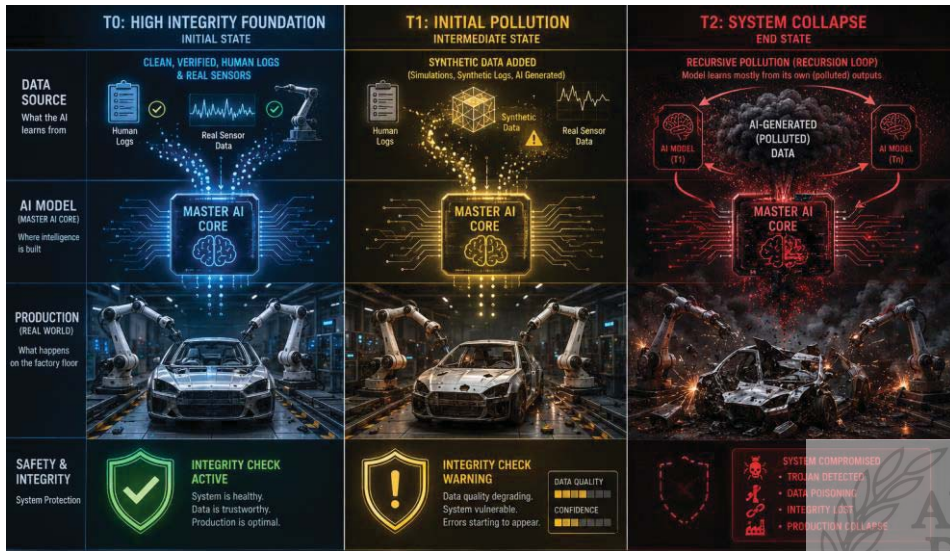


Figure 1. Semantic Data Pollution in Industrial Environment

The earliest manifestations of this problem have been observed in the domain of Large Language Models (LLMs). Modern generative models increasingly train on internet-scale datasets containing massive quantities of publicly available text. However, as AI-generated content proliferates across digital ecosystems – including automatically generated articles, chatbot conversations, synthetic scientific abstracts, and autonomous code generation – the composition of publicly available training data begins to shift. Instead of representing authentic human knowledge production, an increasing portion of the data corpus reflects outputs generated by previous AI systems.

This recursive self-consumption of synthetic information has recently been formalized in machine learning literature under the concept of Model Collapse. Research [3] demonstrated that repeated training on recursively generated synthetic data inevitably leads to progressive degradation of statistical diversity within the model. The process begins with the disappearance of low-probability observations located in the tails of the data distribution. Over successive generations, models increasingly reproduce dominant patterns while gradually forgetting rare events, subtle exceptions, and unusual combinations present in the original training data. In effect, AI systems begin constructing simplified approximations rather than accurate representations of reality.

While the immediate consequences of model collapse in language models may appear relatively benign – manifesting as repetitive text generation, reduced creativity, or factual distortion – the same mechanism becomes substantially more dangerous when transferred into industrial environments. Unlike consumer-facing AI systems, industrial artificial intelligence directly interacts with physical infrastructure, operational decision-making systems, and safety-critical production processes. Errors generated through semantic degradation therefore acquire direct material consequences.

The Industry 4.0 significantly amplifies this challenge, as it is focused primarily on large-scale digitization of production systems through Industrial IoT architectures, cyber-physical systems, cloud computing, and autonomous monitoring infrastructures. In this paradigm, industrial competitiveness depended largely on generating and processing increasingly large quantities of operational data. The assumption underlying the Big Data revolution was straightforward: greater quantities of data would inevitably improve predictive capability.

Industry 5.0 fundamentally changes this assumption. Rather than maximizing automation efficiency alone, Industry 5.0 emphasizes resilience, sustainability, human-centred decision making, and collaborative intelligence between humans and machines. In this new paradigm, the quality and semantic reliability of data become more important than its volume. The growing dependence on synthetic industrial data therefore creates a systemic vulnerability that traditional industrial architectures were never designed to address.

One of the clearest examples emerges through the increasing adoption of digital twins, virtual replicas of physical industrial systems that continuously simulate machine behaviour using real-time operational data. They have become essential tools for predictive maintenance, production optimization, and failure simulation. Modern manufacturing systems increasingly use digital twins to generate synthetic operational scenarios that supplement limited real-world datasets.

At first glance, synthetic data generation through digital twins appears highly beneficial. Rare failure conditions that occur infrequently in physical machines can be artificially simulated thousands of times, allowing predictive maintenance algorithms to train on conditions that would otherwise remain statistically underrepresented. However, the risk emerges when synthetic outputs generated by digital twins gradually become indistinguishable from authentic machine observations.

Consider a predictive maintenance system monitoring industrial turbine vibration patterns in a power generation facility. During early deployment stages, the anomaly detection model trains exclusively on sensor readings collected from real machine operations over several years. Engineers subsequently introduce a digital twin capable of simulating turbine behaviour under various operating conditions. The synthetic data generated by the digital twin is then used to augment training datasets and improve prediction performance.

Over time, the maintenance reporting system itself becomes partially automated through AI-generated incident reports. Future machine learning models are subsequently retrained not only on raw sensor data but also on anomaly labels, maintenance recommendations, and operational summaries generated by earlier AI systems. The training dataset gradually transitions from authentic physical observations toward recursively generated synthetic representations of machine behaviour. At this point, semantic data pollution begins to emerge.

Digital twins inevitably simplify reality. Although simulation models capture dominant operational patterns with high accuracy, they often fail to reproduce highly irregular or unknown machine behaviours. Rare failure modes, manufacturing imperfections, unpredictable environmental interactions, and anomalous combinations of variables remain absent from synthetic simulation environments. Since machine learning systems preferentially learn from statistically dominant patterns, repeated retraining causes models to increasingly adapt toward idealized machine behaviour while losing sensitivity to unexpected anomalies.

A parallel real-world phenomenon has already emerged in autonomous vehicle development. Self-driving systems increasingly train on synthetic driving simulations rather than exclusively on physical road observations (Fig. 2). Although simulation environments successfully reproduce common traffic patterns, system reduces its robustness when autonomous systems encounter highly unusual edge cases absent from synthetic training scenarios. Similar risks could now begin appearing within industrial AI systems.

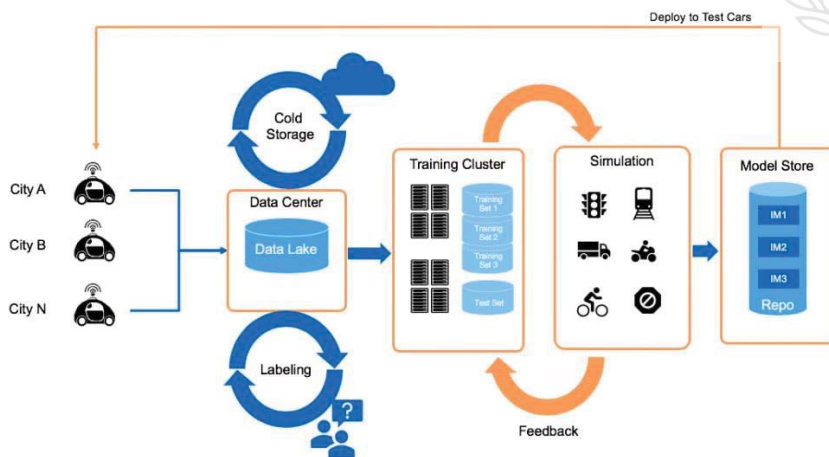


Figure 2. *Data Pipeline for Autonomous Vehicle Development* [9]

The long-term consequences may become severe. A predictive maintenance model operating inside a smart factory may gradually lose its ability to recognize rare precursor signals associated with catastrophic machine failure simply because

those signals are absent from synthetic training datasets generated by digital twins. Since the system continues functioning normally under standard operating conditions, semantic degradation may remain invisible until a rare but critical anomaly occurs.

Future smart factories will likely intensify this recursive dynamic. Emerging industrial architectures increasingly rely on distributed autonomous AI agents capable of independently monitoring machine fleets, optimizing production schedules, generating maintenance documentation, and coordinating supply chain decisions. In such environments, AI systems increasingly exchange machine-generated information with other AI systems in closed feedback loops.

Let us consider a hypothetical future case. In a fully autonomous manufacturing plant in 2035, sensor networks continuously collect machine telemetry data. Digital twins simulate machine degradation trajectories and generate synthetic maintenance scenarios. Predictive maintenance models learn from both physical and simulated operational histories. Automated reporting agents summarize detected anomalies and generate maintenance recommendations. Supply chain optimization systems use these reports to modify inventory planning for spare components. At every stage of this ecosystem, AI-generated semantic artifacts circulate through the industrial information environment. If recursive retraining pipelines are not carefully controlled, future generations of industrial AI may gradually lose direct connection to physical reality while preserving an illusion of operational accuracy.

This introduces a new dimension of cybersecurity risk. Traditional industrial cybersecurity frameworks primarily focus on protecting confidentiality, system integrity, communication security, and unauthorized access prevention [10]. However, semantic data pollution does not originate from malicious intrusion. The system degrades internally through autonomous recursive contamination of its own knowledge structures. In this context, cybersecurity must evolve beyond protecting data infrastructure toward protecting semantic data integrity itself. Future industrial security standards may need mechanisms capable of verifying not only whether data has been altered maliciously, but also whether data still faithfully represents physical reality.

Industry 5.0 places the human operator at the centre of this challenge. Human engineers increasingly serve not merely as supervisors of automated systems, but as essential validators capable of distinguishing genuine physical anomalies from recursively generated synthetic patterns. The concept of Human-in-the-Loop therefore becomes a critical resilience mechanism preventing semantic collapse within industrial AI ecosystems.

Semantic data pollution may ultimately become one of the defining technological risks of Industry 5.0. The challenge is no longer simply protecting industrial systems from external cyberattacks. The emerging challenge lies in preventing intelligent systems from gradually disconnecting themselves from reality through

recursive self-generated knowledge contamination. In future autonomous factories, preserving semantic integrity may become as important as protecting network security itself.

### 3. Cybersecurity Impact Analysis

A critical cybersecurity consequence of semantic data pollution is the gradual loss of anomaly detection capability in AI-driven industrial systems. When machine learning models are repeatedly trained on synthetic or AI-generated data, recursive learning progressively removes statistically rare events located in the tails of the original data distribution. These rare events often represent precisely the abnormal behaviours associated with cyber-attacks, zero-day exploits, sensor manipulation, stealth intrusions, or early-stage machine failures. As recursion amplifies dominant patterns and suppresses low-frequency anomalies, the AI system gradually loses sensitivity to unexpected conditions. Consequently, attacks or mechanical faults that deviate from normal operational patterns may no longer be recognized, creating a dangerous false perception of system stability while vulnerabilities remain undetected.

In modern smart factories, AI models are distributed across Edge–Cloud architectures, where edge devices perform local monitoring while cloud systems aggregate data for centralized learning and model retraining. If semantically polluted or compromised data enters the edge layer, corrupted models may propagate upward to cloud systems during synchronization. Updated cloud models are then redistributed across the factory network, spreading degraded intelligence to all connected machines. A localized anomaly detection failure can therefore cascade system-wide, transforming a single semantic corruption event into a distributed operational and cybersecurity failure across the entire production ecosystem.

To estimate and quantify the semantic data pollution caused by the recursive use of synthetic data, some rigorous statistical and information-theoretic methods can be used. The primary statistical frameworks and metrics used to estimate and track this degradation include optimal transport, distance metrics, information theory, content purity metrics, moment analysis and geometric variance.

To measure how far the synthetic data distribution has drifted from the original, real-world human/industrial baseline distribution across  $n$  generations, distance metrics on probability spaces are highly effective:

- Wasserstein Distance (Earth Mover's Distance): Unlike other divergence metrics, the Wasserstein distance ( $W_2$ ) continuously grows as recursive iterations progress. It measures the minimum effort required to transform the polluted data distribution back into the original baseline distribution. It is exceptionally reliable for tracking long-term, late-stage model collapse [11].

- Kullback-Leibler (KL) Divergence:  $D_{KL}(P||Q)$  measures the relative entropy or information loss when the polluted data distribution (Q) is used to approximate the true industrial baseline distribution (P). While highly sensitive to early-stage distribution shifts, KL divergence can sometimes plateau or behave erratically in late-stage total collapse, making it best suited for early detection [15].

Semantic pollution, which inherently alters the information density of data, can be quantified using information-theoretic criteria:

- Shannon Entropy Tracking: As models recursively train on synthetic data, they tend to drop low-probability edge cases (the long tail) and over-sample the center of the distribution. Tracking the drop in overall dataset entropy quantifies this homogenization [12].
- Perplexity Scaling: By evaluating the recursively trained model against a frozen, high-quality validation set of actual industrial logs or human-curated data, you can measure its perplexity. A monotonic increase in perplexity over training generations directly quantifies how far the model's internal probability distribution has drifted from physical reality.

In industrial settings, data pollution often manifests as a loss of operational variance (e.g., the model forgets what rare equipment failure signatures look like because it only generates nominal operating data) [12].

- Variance Decay ( $\sigma^2$ ) in Gaussian Mixture Models (GMMs): By modelling industrial features as a mixture of distributions, the variance of the components can be tracked statistically. The Gaussian model collapse theorem demonstrates that under uncured recursive loops, the variance of the generated data systematically trends toward zero, eventually reducing the distribution to a single, low-variance peak (a delta function) [13].
- Vocabulary/Token Diversity Coefficients: For text, code, or categorical logs (e.g., SCADA system alarm sequences), the Type-Token Ratio (TTR) or Simpson's Diversity Index can be calculated. A sharp decline in these coefficients across training loops signals that the model is generating repetitive AI slop or formulaic syntax instead of nuanced operational data.

These methods can be combined into a multi-tiered monitoring strategy for setting up an audit pipeline to monitor data pollution in AI-driven industrial processes (Table 1).

*Table 1. Statistical frameworks and metrics for estimating the semantic data pollution*

| Phase of Pollution | Observable Symptom                                  | Best Statistical Tool                                |
|--------------------|-----------------------------------------------------|------------------------------------------------------|
| Early Stage        | Loss of minority data / rare industrial edge cases. | KL Divergence and Shannon Entropy                    |
| Mid Stage          | Narrowing of outputs; systemic semantic drift.      | Perplexity Scaling and Type-Token Ratios             |
| Late Stage         | Homogenized, low-variance, or nonsensical data.     | Wasserstein Distance ( $W_2$ ) and Variance Tracking |

#### 4. Proposed Solutions

In AI-driven industrial processes – ranging from automated SCADA log interpretation to automated predictive maintenance scheduling – the recursive training of models on synthetic or model-generated data induces systemic degradation. This phenomenon, academically formalized as Model Collapse or the Entropy Spiral, manifests when an estimator  $M_{n+1}$  is trained on the output distribution  $p_n(y|x)$  of its predecessor  $M_n$ , rather than the true underlying physical or operational distribution  $p^*(y|x)$ .

Statistically, if the true distribution possesses a dense support, the sampling noise and bounded functional capacity of  $M_n$  cause it to misapproximate the tails of  $p^*$ . Over successive generations  $n \rightarrow \infty$ , the conditional variance decreases ( $\sigma_{n+1}^2 < \sigma_n^2$ ), shifting the probability mass toward a singular, low-variance mode. In industrial environments, this means rare but critical anomalies (e.g., specific vibration signatures preceding a turbine failure) are treated as statistical noise and eliminated from the synthetic training corpus.

To mitigate this semantic data pollution, three core architectural paradigms can be used: Cryptographic Data Provenance, Statistical and Algorithmic Watermarking, and Human-in-the-Loop (HITL) Curation Frameworks.

##### 4.1. Data Provenance

One of the most fundamental defences against semantic data pollution is maintaining strict provenance tracking for all data entering industrial machine learning pipelines. Data provenance refers to the ability to trace the origin, transformation history, ownership, and processing path of data throughout its entire lifecycle.

In industrial environments, machine learning systems increasingly integrate heterogeneous data streams originating from physical sensors, enterprise systems, digital twins, simulation environments, maintenance logs, and AI-generated anomaly reports. Without provenance metadata, future retraining systems cannot distinguish authentic observations from synthetic or recursively generated artifacts.

A possible architectural solution involves attaching immutable metadata descriptors to every data instance:

$$D_i = \{x_i, S_i, T_i, C_i, A_i\} \quad (1)$$

where:

$x_i$  = observed data sample

$S_i$  = source identifier (sensor, simulation, AI-generated output)

$T_i$  = timestamp

$C_i$  = confidence score

$A_i$  = transformation history

This framework allows weighting data differently during retraining procedures. For example, predictive maintenance systems could define weighted learning functions:

$$\sum_{i=1}^n \frac{w_i}{l(x_i)} \quad (2)$$

where provenance-based weighting coefficient ( $w_i$ ) penalizes synthetic or recursively generated samples.

A practical implementation may involve blockchain-based industrial audit trails, although current evidence suggests blockchain introduces scalability limitations in high-frequency industrial sensor environments. Lightweight distributed ledgers or signed metadata registries may prove more practical. The limitation of provenance tracking lies in its inability to assess semantic correctness directly. Provenance identifies origin, but authentic sensor data itself may still contain drift, corruption, or adversarial manipulation.

#### 4.2. Watermarking

A second mitigation strategy involves explicit identification of synthetic data through watermarking mechanisms.

As AI systems increasingly generate industrial reports, simulation outputs, maintenance recommendations, and synthetic operational datasets, future retraining pipelines require mechanisms capable of distinguishing machine-generated content from authentic physical observations.

Digital watermarking has already emerged in generative AI research as a method for identifying synthetic text, images, and generated media. Similar concepts may be extended into industrial machine learning environments. A simple statistical watermark can be embedded by intentionally introducing traceable distribution signatures into synthetic data. Let synthetic data distribution be represented by  $P_s(x)$ . A watermark function introduces controlled perturbation:

$$P'_s(x) = P_s(x) + \varepsilon(x) \quad (3)$$

where  $\varepsilon(x)$  represents a statistically detectable signature.

During retraining, statistical hypothesis testing can evaluate contamination probability:

$$H_0: P(x) = P_r(x) \quad (4)$$

$$H_1: P(x) = P'_s(x) \quad (5)$$

using divergence metrics such as Kullback-Leibler divergence, Jensen-Shannon divergence or Wasserstein distance [14].

For example, a digital twin generating synthetic machine vibration signals may embed small statistical signatures within non-critical frequency bands. Future retraining systems could automatically reject heavily synthetic datasets once contamination thresholds exceed predefined limits.

However, watermarking remains experimentally immature. Adversarial optimization methods may eventually remove statistical signatures, while repeated data transformations can weaken watermark detectability. At present, watermarking should be viewed as a probabilistic detection layer rather than a complete solution.

#### 4.3. Human-in-the-loop

Industry 5.0 fundamentally repositions human operators as active participants in intelligent industrial ecosystems rather than passive supervisors. This principle becomes critically important in preventing semantic data pollution. Unlike automated systems, human experts possess contextual understanding of rare operational anomalies, unexpected environmental conditions, and physically implausible machine behaviours that statistical models frequently overlook.

The central problem of recursive learning is that low-frequency edge cases gradually disappear from training distributions. Human intervention becomes essential precisely because humans can identify semantically meaningful anomalies independent of statistical frequency.

A hybrid architecture can be modeled as:

$$\text{Decision} = \alpha \cdot \text{AI}(x) + (1 - \alpha) \cdot \text{Human}(x) \quad (6)$$

where:

$x$  = the input data vector

$\text{AI}(x)$  = machine inference: conditional probability or scoring vector output

$\text{Human}(x)$  = expert semantic validation score

$\alpha$  = automation confidence parameter (trust weight)

In high-risk operational states, the system dynamically reduces automation dominance  $\alpha \rightarrow 0$ , forcing human verification.

A practical example exists in modern aircraft predictive maintenance systems. Automated anomaly detection may flag unusual vibration signatures, but final maintenance decisions still require certified engineers because rare structural fatigue events often remain underrepresented in historical datasets. Smart factories of the future may require similar architectures in which human operators periodically validate retraining datasets before model deployment.

The limitation of this approach is scalability. Human supervision introduces cost, latency, and dependence on expert availability, particularly in fully autonomous manufacturing environments.

## 5. Deployment

### 5.1. Implementation challenges, operational constraints and associated costs

Implementing defences against semantic data pollution within a live smart factory environment introduces severe technical bottlenecks, such as legacy infrastructure, Edge-Cloud asymmetry and the dimensionality.

Modern factories rely heavily on legacy programmable logic controllers (PLCs) and Supervisory Control and Data Acquisition (SCADA) networks that operate under strict real-time, low-latency execution constraints (often down to less than 10 ms). Injecting a cryptographic metadata tuple at the sensor or edge layer requires computational overhead that standard PLCs lack the memory or processing power to compute. Embedding localized statistical perturbations into generated control data streams risks destabilizing tight closed-loop control systems if the signature inadvertently excites unmodeled mechanical resonance frequencies of physical equipment.

Industrial data pipelines frequently employ aggressive data compression, downsampling, and dimensionality reduction techniques to manage high-frequency sensor streams. These lossy compression layers often treat low-amplitude statistical watermarks or high-frequency orthogonal signatures as random noise, systematically stripping them out during ingestion. Consequently, watermarking becomes inherently brittle when data transitions through multi-vendor middleware layers.

Finally, deploying a multi-layered semantic security architecture introduces structural costs, which include computational overhead (if blockchain is used to verify data origins), operation latency, and human capital expenditure needed for specialized industrial experts with deep industrial process knowledge.

## 5.2. Illustrative industrial scenario and quantitative assessment

Consider a high-throughput steel rolling mill utilizing automated AI models to monitor bearing degradation trajectories in heavy-duty rollers based on high-frequency vibration telemetry.

- **Baseline Regime:** The plant accumulates real sensor data over several years. Due to the high reliability of the equipment, critical bearing failures occur only once every 18 months (representing the "heavy tail" of the true distribution  $p^*(x)$ ).
- **The Pollution Event:** To improve predictive capabilities for rare faults, plant operators introduce a high-fidelity Digital Twin. The Digital Twin generates synthetic vibration profiles to expand the training set. Over 5 retraining generations ( $n = 5$ ), the data pipeline suffers from uncured recursive learning as the model begins consuming its own automated maintenance reports and synthetic logs.

Let the true bearing vibration anomaly signature be characterized by a multi-dimensional feature variance  $\sigma_*^2$ . Under the recursive uncured training loop, we model the contraction of the estimated variance  $\sigma_n^2$  across generations using a simplified variance decay coefficient  $\gamma$ , influenced by the proportion of synthetic data leakage  $\lambda$ :

$$\sigma_n^2 = \sigma_{n-1}^2 \cdot (1 - \lambda \cdot (1 - \gamma)) \quad (7)$$

Assuming a synthetic leakage rate of  $\lambda = 0,60$  (60% synthetic data in the retraining pool) and an approximation loss of  $\gamma = 0,85$ , after 5 generations, the operational variance shrinks systematically:

$$\sigma_5^2 = \sigma_0^2 \cdot (0,91)^5 \approx 0,624 \cdot \sigma_0^2 \quad (7)$$

The model loses approximately 37.6% of its original statistical variance, shifting its probability mass tightly around the nominal, healthy operating mode.

Without defence layer, when a real, physical micro-fracture begins to manifest in a roller bearing, its unique vibration signature falls outside the narrow, collapsed distribution peak of  $M_5$ . The AI misclassifies the warning as nominal signal noise. The failure propagates unmitigated, resulting in an unpredicted structural bearing breakdown. This leads to 36 hours of unscheduled plant downtime, costing an estimated 180.000 € in lost production and hardware replacement.

With multi-layer semantic security implemented, the active monitoring pipeline tracks Shannon Entropy and Wasserstein Distance ( $W_2$ ). By generation  $n = 3$ , the pipeline flags a  $W_2$  divergence surge exceeding the pre-established security threshold, triggering an automated integrity warning. The system then calculates a high epistemic uncertainty score for the drifting data segments, driving the

automation confidence parameter  $\alpha \rightarrow 0,15$ . This dynamically routes the flagged data assets to a human-in-the-loop validation queue. A certified maintenance engineer reviews the anomaly, manually isolates the synthetic digital twin outputs, and restores the empirical baseline training data.

## 6. Conclusion

The increasing integration of artificial intelligence into autonomous industrial environments introduces a previously underrecognized systemic vulnerability arising from recursive dependence on synthetic data. This paper proposed the concept of Semantic Data Pollution as a new failure mechanism in AI-driven manufacturing systems, where repeated retraining on machine-generated outputs gradually disconnects intelligent systems from physical reality. Building upon recent theoretical findings on model collapse in generative AI, we demonstrated how recursive learning progressively suppresses statistically rare but operationally critical anomalies, thereby reducing the effectiveness of predictive maintenance, anomaly detection, and cybersecurity monitoring systems.

Unlike conventional cybersecurity threats that originate from malicious external actors, semantic data pollution represents an endogenous degradation process emerging from the internal dynamics of intelligent systems themselves. In distributed Edge–Cloud industrial architectures, localized semantic corruption may propagate across entire production ecosystems, creating large-scale operational vulnerability despite the absence of direct cyber intrusion.

The transition toward Industry 5.0 fundamentally changes the role of cybersecurity. Protecting industrial systems will no longer involve only safeguarding data confidentiality, integrity, and availability, but also preserving the semantic fidelity of data used for autonomous machine learning. Mitigation strategies such as provenance-aware training pipelines, synthetic data watermarking, statistical drift monitoring, and Human-in-the-Loop validation provide promising directions, although none currently offers complete protection.

Future research should focus on formal mathematical modelling of semantic degradation dynamics and experimental validation using real industrial AI systems. Semantic integrity may ultimately become a foundational requirement for trustworthy autonomous factories, requiring the development of new cybersecurity standards specifically designed for AI-native industrial environments.

## References

- [1] Lemeš, S. (2024). Informaciona sigurnost u Industriji 4.0. in *Industrija 4.0: digitalna transformacija oblikuje budućnost* (Karabegović I. i Majstorović V., editors), ISBN 978-9958-9262-8-0, 978-9926-410-94-0 Robotics Society of Bosnia and Herzegovina; Academy of Sciences and Arts of Bosnia and Herzegovina, pp 501-524
- [2] Cléménçon, S., Sabourin, A. (2026). Weak signals and heavy tails: learning theory meets extreme value analysis. *Extremes* 29, 303–345 <https://doi.org/10.1007/s10687-026-00540-z>
- [3] Shumailov, I., Shumaylov, Z., Zhao, Y. et al. (2024). AI models collapse when trained on recursively generated data. *Nature* 631, 755–759 <https://doi.org/10.1038/s41586-024-07566-y>
- [4] Seddik, M. E. A., Chen, S. W., Hayou, S., Youssef, P., & Debbah, M. (2024). How bad is training on synthetic data? a statistical analysis of language model collapse. *arXiv preprint* <https://doi.org/10.48550/arXiv.2404.05090>
- [5] Martínez-Arellano, G., Ratchev, S. (2024). Improving the Development and Reusability of Industrial AI Through Semantic Models. In: Thiede, S., Lutters, E. (eds) *Learning Factories of the Future. CFL 2024. Lecture Notes in Networks and Systems*, vol 1059. Springer, Cham. [https://doi.org/10.1007/978-3-031-65411-4\\_22](https://doi.org/10.1007/978-3-031-65411-4_22)
- [6] Li, X., Cheng, Y., Möller, C., & Lee, J. (2025). Data issues in industrial AI systems: A meta-review and research strategy. *Computers in Industry*, 173, 104361. <https://doi.org/10.1016/j.compind.2025.104361>
- [7] Hu, Z., Rostami, M., & Thomason, J. (2025). Multi-modal Synthetic Data Training and Model Collapse: Insights from VLMs and Diffusion Models. *arXiv preprint* <https://doi.org/10.48550/arXiv.2505.08803>
- [8] ISO/IEC FDIS 27090 Cybersecurity — Artificial Intelligence — Addressing security threats and compromises to artificial intelligence systems
- [9] Rao, S. (2019). Autonomous Driving Challenges, How to Build a Data Pipeline for Autonomous Driving, <https://www.netapp.com/blog/how-to-build-a-data-pipeline-for-autonomous-driving/> (accessed 22.6.2026)
- [10] ISO/IEC 24392:2023 Cybersecurity — Security reference model for industrial internet platform (SRM- IIP)
- [11] Borji, A. (2024). A Note on Shumailov et al. (2024): AI Models Collapse When Trained on Recursively Generated Data'. *arXiv preprint* <https://doi.org/10.48550/arXiv.2410.12954>
- [12] Vu, H. A., Reeves, G., & Wenger, E. (2025). What happens when generative AI models train recursively on each others' outputs?. *arXiv preprint* <https://doi.org/10.48550/arXiv.2505.21677>

- [13] Wondrasek , J. A. (2026). The Entropy Spiral: How AI Slop Degrades Future AI Systems Through Recursive Training, <https://www.softwareseni.com/the-entropy-spiral-how-ai-slop-degrades-future-ai-systems-through-recursive-training/> (accessed 22.6.2026)
- [14] Piryanov, K., Grigorova, I., Karamfilov, A., & Efremov, A. (2026). Automated Data Monitoring Using a Canberra-Based Drift Score. Applied Sciences, 16(5), 2232. <https://doi.org/10.3390/app16052232>
- [15] GeeksforGeeks (2025). Kullback Leibler (KL) Divergence. <https://www.geeksforgeeks.org/machine-learning/kullback-leibler-divergence/> (accessed 22.6.2026)

