

When AI Starts Training on AI: Model Collapse, AI Slop, and the Emerging Crisis of Data Provenance

Snigdha Singh Dewal

Ph.D Scholar at UPES, Dehradun, India

DOI: <https://doi.org/10.51244/IJRSI.2026.1308000186>

Received: 24 August 2026; Accepted: 02 September 2026; Published: 14 September 2026

ABSTRACT

Generative artificial intelligence systems no longer merely consume the human-authored internet; they now produce a substantial share of it. As synthetic text, images and data circulate back into the corpora used to train successive model generations, researchers have identified a degenerative process termed "model collapse," in which recursive training on machine-generated outputs causes models to progressively lose information about the rare, minority and low-frequency features of the original data distribution. This paper distinguishes model collapse, a training-dynamics phenomenon, from the related but distinct problem of "AI slop," a content-quality phenomenon, and argues that their interaction produces a more consequential legal concern: the erosion of data provenance. Drawing on the foundational 2024 Nature study and its subsequent refinements, together with regulatory developments such as the EU AI Act's training-data transparency obligations and technical standards including the Coalition for Content Provenance and Authenticity framework, the paper develops the concept of "epistemic due diligence" as an emerging obligation for AI developers and regulators alike. It concludes that the central resource constraint on AI development is shifting from computational power and raw data volume toward the scarcer commodity of verifiably authentic, traceable and diverse information, and that law is only beginning to develop the doctrinal tools required to govern this shift.

Keywords Model Collapse; Synthetic Data; Data Provenance; AI Governance; EU AI Act; Epistemic Due Diligence

INTRODUCTION AND LITERATURE REVIEW

The internet was, until recently, a predominantly human artefact. Articles, research papers, photographs, reviews, forum posts and code repositories were, overwhelmingly, records of human activity, and it was this enormous archive that early generations of large language models were trained upon. That relationship is now reversing. Generative systems increasingly author a meaningful share of newly published online material, ranging from carefully edited long-form writing to unreviewed, mass-produced text, images and video. Once published, such material becomes largely indistinguishable from human-authored content, and it is this indistinguishability, rather than the mere existence of synthetic content, that gives rise to the central concern examined in this paper: an emerging feedback loop in which human-generated data trains AI systems, whose outputs are published online and subsequently scraped to train the next generation of models.

The technical literature has given this feedback loop a name. Shumailov et al. (2024), in a widely cited Nature study, demonstrated that indiscriminate recursive training on model-generated content produces a degenerative process they termed model collapse. The study identified two stages: an early stage in which models lose information residing in the statistical "tails" of a distribution rare events, minority viewpoints, unusual phrasing and low-frequency observations and a late stage in which the model's output distribution converges toward a substantially narrower approximation of reality. The unsettling feature of this process is that it need not manifest as an obvious failure. A collapsing model can remain fluent, score adequately on conventional benchmarks, and still become progressively less diverse and less representative of the world it was meant to describe.

It is important, however, to distinguish model collapse from the popular concept of "AI slop," a term used to describe low-effort, low-quality synthetic content circulating online. Model collapse is a training-dynamics problem; AI slop is an information-quality problem. The two nonetheless interact in a manner with particular relevance to law: where multiple websites republish or paraphrase the same AI-generated proposition, a subsequent model scraping that material may encounter what appears to be independent corroboration but is, in substance, a single error reproduced under different bylines. This paper terms this phenomenon "synthetic consensus," and situates it as a central mechanism by which factual and legal inaccuracies could become entrenched in future training corpora.

A second strand of literature addresses the scarcity of human-generated data that makes synthetic data attractive in the first place. Villalobos et al. (2024), in research presented at ICML, estimated the effective stock of quality-adjusted public human-generated text at approximately 300 trillion tokens, projecting that frontier models could approach comparable data volumes between 2026 and 2032. This projected scarcity has driven developers toward synthetic data as a scalable substitute, raising the question addressed by a third strand of literature: whether synthetic data is inherently degrading. Gerstgrasser et al. (2024) and the subsequent Kazdan et al. (2025) study found that while replacing real data entirely with successive generations of synthetic data produces collapse, accumulating synthetic data alongside retained authentic data can avoid the same dynamic. Kang et al. (2025), examining large-scale pre-training, similarly reported mixed results depending on whether synthetic data was rephrased from authentic sources or generated *de novo*, concluding that a properly curated mixture of synthetic and natural data can be beneficial rather than harmful.

A further, more recent contribution comes from Jangjoo et al. (2026), published in *Physical Review Letters*, which found in analytically tractable statistical models rather than frontier-scale systems that introducing even a single external, non-model-generated data point into an otherwise closed training loop could prevent collapse. While the authors caution against over-extrapolating this result to large language models, the conceptual insight is significant for governance: a closed informational loop behaves fundamentally differently from a system that retains access to information generated outside itself. Parallel work has begun to examine whether collapse dynamics also apply to fairness and bias, with emerging 2026 research suggesting that demographic bias may amplify recursively and that fairness degradation can become detectable before conventional performance metrics show measurable decline.

Finally, the regulatory and standard-setting literature has begun to respond to these technical findings. The European Commission's implementation of Article 53 of the EU AI Act now requires providers of general-purpose AI models to publish a summary of training content, including whether that content was scraped, licensed, user-derived or synthetic, an obligation that became applicable from 2 August 2025. The U.S. Copyright Office's 2025 Part 3 report on generative AI training addressed data acquisition, licensing and fair use, though largely without reference to the recursive-training problem. Technical standard-setting bodies, principally the Coalition for Content Provenance and Authenticity and the U.S. National Institute of Standards and Technology, have meanwhile developed cryptographically verifiable provenance architectures capable of recording a content object's creation and editing history, including whether AI systems were involved at any stage. This paper synthesises these three literatures—the computer science of model collapse, the economics of data scarcity, and the emerging law of training-data transparency—to argue that the legal question is shifting from data ownership toward data integrity.

METHODOLOGY

This paper adopts a doctrinal and interdisciplinary analytical method, combining a structured review of the technical computer science literature on model collapse with a comparative analysis of emerging regulatory and technical-standard instruments governing training-data transparency. The technical literature was selected on the basis of three criteria: methodological rigour (peer-reviewed publication in venues including *Nature*, *Physical Review Letters*, *ICML/PMLR* and *EMNLP*, or, where the research is very recent, reputable pre-publication venues clearly identified as such), direct relevance to recursive training dynamics, and temporal currency, with priority given to studies published between 2024 and 2026 to capture the rapid evolution of this specific research question.

The regulatory and policy analysis draws primarily on three categories of primary source: binding legal instruments (the EU AI Act and its Article 53 implementing template on general-purpose AI training-content transparency), quasi-legislative reports (the U.S. Copyright Office's Part 3 report on generative AI training), and technical standards developed by multi-stakeholder bodies (the C2PA Content Credentials specifications and NIST's guidance on synthetic-content risk mitigation). These instruments were chosen because they represent, respectively, the most advanced binding transparency obligation currently in force, the most comprehensive contemporary study of the copyright dimension of AI training, and the leading technical architecture for recording content provenance.

The analytical method proceeds in three stages. First, the paper maps the causal architecture of model collapse as established in the computer science literature onto categories familiar to legal analysis—namely, provenance, authenticity, chain of custody and accountability—in order to translate a technical phenomenon into a set of governable legal questions. Second, it examines the extent to which existing regulatory instruments already address these questions, identifying gaps between what the technical literature suggests is necessary (for example, the preservation of authentic data as an anchor against recursive drift) and what current law actually requires. Third, it derives a set of proportionate governance recommendations addressed separately to AI developers and to regulators, informed by a risk-based approach that calibrates obligations to the systemic significance of the AI system in question.

This paper does not present original empirical or experimental findings; its contribution is synthetic and doctrinal, situating recent and fast-moving technical findings within an emerging legal framework. Limitations of this method include the rapidly evolving nature of the underlying computer science literature, several components of which remain pre-publication or early-stage at the time of writing, and the corresponding risk that specific empirical estimates (for example, projected data-exhaustion timelines) may be revised as further research is published. The analysis therefore emphasises structural and conceptual conclusions—the distinction between model collapse and AI slop, the mechanism of synthetic consensus, and the proposed standard of epistemic due diligence—over any single numerical projection.

RESULTS AND DISCUSSION

The central finding of this analysis is that model collapse, properly understood, is not primarily a prediction that artificial intelligence will become unusable. The literature reviewed above is consistent in finding that synthetic data is not inherently degrading and that collapse is neither inevitable nor, on current evidence, already underway across frontier systems. Gerstgrasser et al. (2024) and Kazdan et al. (2025) in particular demonstrate that accumulating synthetic data alongside retained authentic data, rather than substituting one for the other, can avoid collapse dynamics altogether. The more defensible and, from a legal perspective, more consequential finding is narrower: uncontrolled, indiscriminate recursive dependence on unlabelled synthetic data creates a material risk of information degradation, and that risk is presently unmanaged by law. This reframing has three legal implications. The first concerns synthetic consensus. Where AI-generated propositions are republished, paraphrased and indexed across many ostensibly independent sources, a subsequent model—or, equally, a human researcher, journalist or lawyer conducting online research—may mistake repetition for corroboration. The risk is particularly acute in fields where factual precision carries legal consequences, including case citation, statutory interpretation, medical guidance and financial disclosure. Existing legal doctrines addressing misinformation, defamation and professional negligence were not designed with this dynamic in mind, since they generally presuppose an identifiable originating publisher rather than a distributed, self-reinforcing chain of derivative machine-generated statements.

The second implication concerns the object of transparency regulation. The EU AI Act's Article 53 training-content transparency template is significant precisely because it requires disclosure not merely of whether content is AI-generated, but of what categories of material—publicly scraped data, private datasets, user data and synthetic data—were used to train a model, and requires that disclosure to be updated where subsequent training materially changes its composition. This represents a meaningful advance over disclosure regimes that focus solely on the labelling of AI-generated outputs, since it addresses the input side of the pipeline where the risk of recursive contamination actually accumulates. However, the obligation currently applies

only to general-purpose AI models and stops short of requiring developers to demonstrate active governance of synthetic-data accumulation, contamination auditing, or preservation of authentic data anchors precisely the interventions the computer science literature identifies as effective. There is, accordingly, a gap between the technical measures shown to mitigate collapse and the substantive content of the leading transparency obligation currently in force.

The third implication concerns provenance architecture more broadly. Technical frameworks such as C2PA's Content Credentials demonstrate that provenance can, in principle, be recorded as a chain of verifiable claims human-created, AI-assisted, human-edited, AI-translated, human-approved rather than as a binary human/AI label, which is important given that watermarks can be stripped, metadata removed, and content freely copied, edited or translated across systems. Yet provenance-by-design raises a legitimate countervailing concern: a persistent, granular record of a work's creation and editing history could itself become a surveillance mechanism, exposing authorship, editing behaviour, location metadata or proprietary workflows. The paper's analysis suggests that governance frameworks should therefore pursue privacy-preserving provenance verifiable at the level required for accountability without constituting a universal, publicly accessible ledger of individual creative or professional activity a balance that C2PA's design principles gesture toward but that binding law has not yet firmly settled.

Taken together, these findings support the paper's central conceptual proposal: that AI governance is moving, or should move, from asking who owns a given dataset toward asking whether that dataset can be trusted, a standard this paper terms epistemic due diligence. Operationally, epistemic due diligence would require an AI developer to be able to answer, for any given training corpus, where the information originated, whether it is independently corroborated or is instead a repeated instance of the same synthetic proposition, whether minority and low-frequency observations are being systematically lost across training generations, and whether output distributions have measurably drifted over time. None of these questions map neatly onto existing categories of intellectual property or data protection law, which is precisely why the paper argues for their treatment as a distinct, emerging doctrinal category rather than an extension of existing frameworks.

The discussion also identifies a practical, risk-based path forward for both developers and regulators, rather than a blanket prohibition on synthetic training data, which the reviewed literature suggests would be both scientifically unwarranted and practically unworkable given the projected scarcity of fresh human-generated text identified by Villalobos et al. (2024). For developers, this includes maintaining provenance-aware datasets, segregating and labelling synthetic data rather than allowing it to become indistinguishable from authentic material, preserving verified real-data anchors across training generations, and subjecting high-risk domains such as law, medicine and finance to heightened verification. For regulators, this includes extending meaningful training-data transparency obligations proportionate to systemic risk, encouraging interoperable provenance standards such as C2PA, and establishing independent auditability for high-impact systems, while calibrating the intensity of obligations to the severity of the harm a given AI application could cause if trained on contaminated or synthetically homogenised data.

CONCLUSION

This paper has argued that the popular narrative of AI "cannibalising" itself into uselessness overstates what the current technical literature actually supports. Model collapse is a real and well-documented phenomenon under specific recursive training conditions, but it is neither inevitable nor, on the evidence reviewed, already unfolding at the level of frontier systems; synthetic data, used with appropriate curation and alongside retained authentic data, can be beneficial rather than degrading. The more accurate and more legally significant concern is narrower and more insidious: an internet in which synthetic and human-generated material become increasingly difficult to distinguish, in which repeated machine-generated propositions can masquerade as independent corroboration, and in which the rare, minority and dissenting information that constitutes a healthy information ecosystem is gradually, and not always visibly, lost.

The paper has further argued that this shift reframes the central question in AI law. Where earlier debates asked who owns the data used to train a model, and more recent debates ask whether such use is lawful, the

emerging and, this paper suggests, more consequential question is whether the data can be trusted at all a standard proposed here as epistemic due diligence. Existing regulatory developments, particularly the EU AI Act's Article 53 training-content transparency obligation and the technical provenance architecture developed by C2PA and supported by NIST, represent meaningful but partial progress toward this standard. Neither yet requires developers to demonstrate active governance against recursive contamination, synthetic consensus, or the loss of distributional diversity identified in the computer science literature as the substantive mechanism of collapse.

Future research and regulatory attention should focus on three priorities. First, the development of proportionate, risk-calibrated obligations requiring high-risk and systemic-risk AI developers to document not merely the volume but the composition and provenance of training data, including explicit treatment of synthetic content. Second, continued investment in interoperable, privacy-preserving provenance infrastructure that avoids converting content authentication into a tool of surveillance. Third, sustained public and institutional investment in preserving high-quality, verified, human-generated and primary-source datasets, not because such data is scarce in an absolute sense, but because it functions as the epistemic anchor without which recursive training risks compounding rather than correcting its own errors. The central resource constraint on the future of artificial intelligence, this paper concludes, is unlikely to remain computational power alone; it is increasingly likely to be trusted, traceable and diverse information infrastructure that, unlike compute, cannot simply be manufactured on demand.

REFERENCES

1. Coalition for Content Provenance and Authenticity (C2PA). (2024). Content Credentials and AI/ML provenance specifications. C2PA.
2. European Commission. (2025). General-purpose AI training content transparency template (implementing Article 53 of the EU AI Act). European Commission.
3. Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., Korbak, T., Agrawal, R., Pai, D., Gromov, A., Roberts, D. A., Yang, D., Donoho, D. L., & Koyejo, S. (2024). Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. arXiv preprint.
4. Jangjoo, N., et al. (2026). Lost in retraining: Closed-loop learning and model collapse in exponential families. *Physical Review Letters*.
5. Kang, S., et al. (2025). Demystifying synthetic data in LLM pre-training. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
6. Kazdan, J., Schaeffer, R., Dey, A., Gerstgrasser, M., Rafailov, R., Donoho, D. L., & Koyejo, S. (2025). Collapse or thrive: Perils and promises of synthetic data in a self-generating world. *Proceedings of the 42nd International Conference on Machine Learning (ICML/PMLR)*.
7. National Institute of Standards and Technology (NIST). (2024). Reducing risks posed by synthetic content: An overview of technical approaches to digital content transparency. U.S. Department of Commerce.
8. Shabgahi, S., et al. (2026). ForTIFAI: Fending off recursive training induced failure for AI model collapse. *npj Artificial Intelligence*.
9. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759.
10. U.S. Copyright Office. (2025). Copyright and artificial intelligence, Part 3: Generative AI training (pre-publication version). Library of Congress.
11. Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). Will we run out of data? Limits of LLM scaling based on human-generated data. *Proceedings of the 41st International Conference on Machine Learning (ICML/PMLR)*.
12. Ahrefs. (2025). Analysis of AI-generated content prevalence in newly created web pages, April 2025. Ahrefs.