Memory, Mind & Media (2026), 5, e13, 1–12
doi:10.1017/mem.2026.10039

Memory, Mind & Media

SHORT RESEARCH ARTICLE

Ghost in the cache: How data decay shapes the unseen landscape of AI memory

Hanna Gaweł

Institute of Information Studies, Jagiellonian University, Kraków, Poland
Email: hanna.gawel@uj.edu.pl

(Received 29 August 2025; revised 06 May 2026; accepted 17 May 2026)

Abstract

The assumption that deleted digital information is irretrievably lost remains widespread, yet this belief obscures the continued existence of residual traces within archives, backups, and systems that implement only partial or 'soft' deletion. Such remnants, though frequently overlooked, may persist in shaping the operation of artificial intelligence (AI) and therefore warrant critical examination. This article addresses three interrelated questions. First, through what mechanisms do obsolete or concealed data fragments re-enter AI systems and exert influence that manifests as bias or distortion? Second, in what ways might the trajectory of digital information be conceptualised through reference to ecological processes such as decomposition, renewal, and systemic adaptation? Third, what forms of regulatory or procedural innovation – here articulated as a model of 'digital composting' – could facilitate the identification, evaluation, and responsible management of residual data? The analysis demonstrates that outdated data frequently function as 'ghost inputs'. Despite their invisibility, these elements shape the generative capacity of AI, modify the narratives it produces, and subtly recalibrate public discourse as well as shared cultural memory. Their persistence underscores the communicative and social significance of digital traces once presumed to be obsolete. To advance this discussion, the article introduces the concept of a 'Data Decay Pathway'. This framework offers a novel means of addressing accountability and transparency in digital systems, emphasising that processes of digital decay, much like those observed in natural ecosystems, may serve either to sustain the vitality of collective memory or to propagate forms of distortion and systemic vulnerability.

Keywords: information management; data decay; ghost inputs; information ecology; AI memory

In an era obsessed with remembering everything, we often overlook what happens to data we think have been 'deleted'. This article looks at how old, forgotten digital information – like records that were 'soft-deleted' or stored away in archives – can still exist and influence artificial intelligence (AI) systems.

Using ideas from how natural ecosystems work and how memory functions, this study answers three key questions:

- How do these old data fragments get back into AI models and cause them to act in biased ways?

- How can we use the idea of natural decomposition to understand the life of digital information?
- What new rules or systems, like a 'digital compost' protocol, could help us find, check, and properly handle this obsolete data?

My research found that these old data footprints, which we thought were harmless, act like 'ghost inputs'. They strengthen hidden biases and change the stories that AI creates. From a communication and media perspective, the silent survival of these old records can influence our shared memory. Generative AI systems piece together these broken pieces into new stories, subtly changing public conversations and how we make sense of things. Just like in a natural ecosystem, this digital decay can either help or harm the health of our AI memory systems.

This study introduces the 'Data Decay Pathway' as a new concept to help us talk about accountability and transparency in the digital world. By seeing data decay as a useful resource instead of just a loss, this work connects information science, media studies, and AI ethics. The 'data-compost' idea offers practical steps for developers and policymakers to manage the hidden life of erased data, leading to more responsible and sustainable AI systems.

Introduction

Historically, human and societal memory relied upon the 'virtue of forgetting' (Mayer-Schonberger 2011), a natural degradation that cleared space for new synthesis. However, contemporary discourse routinely assumes that digital environments afford near-comprehensive preservation, functionally suspending this natural mortality and initiating what Hoskins (2013) terms the 'end of decay time'. Vast newsfeeds and ubiquitous cloud backups are commonly cited as evidence of an unprecedented archival capacity. This narrative of comprehensive capture, however, occludes a less conspicuous but consequential phenomenon – the continued existence and influence of data thought to have been expunged. Soft-deleted records, orphaned logs, and materials consigned to cold storage do not simply disappear; they persist as 'ghost inputs' within the infrastructural substratum of networked systems (Thylstrup 2014). Within this digital 'forgetting ecology', data lose their original context but resist total decomposition. Drawing on an ecological metaphor, this article contends that scholarly attention fixated on total recall has rendered invisible an essential dimension of data decay: an unattended layer of informational residue that materially shapes the memories constructed by computational systems and, by extension, collective recollection (Nardi and O'Day 1999; Floridi 2010).

There are both practical and conceptual reasons to attend to this 'quiet rot'. Empirically, the epistemic and predictive capacities of machine-learning systems depend crucially on the provenance and quality of their inputs. Unmanaged decay can permit obsolete records to re-enter training pipelines, reproducing historical inaccuracies and amplifying latent biases. Conceptually, the phenomenon extends beyond mere technical dysfunction. From a communication and media studies vantage, the afterlives of deprecated data participate in the reconfiguration of narrative, as generative systems recombine fragmentary inputs into emergent storylines, thereby exerting subtle influence on public discourse and collective meaning-making (Barnier and Hoskins 2018; Syvertsen *et al.* 2019). Unattended decay, therefore, functions both as a vector of algorithmic distortion and as an active contributor to the ongoing remaking of mediated pasts.

To interrogate these dynamics, this article pursues three interrelated and deliberately synthetic questions:

https://doi.org/10.1017/mem.2026.10039 Published online by Cambridge University Press

(1) By what mechanisms do decayed data artefacts re-enter generative models and affect their outputs?
(2) How might analogies from natural decomposition illuminate the lifecycle of digital traces within informational ecosystems?
(3) What governance and design interventions – provisionally framed as 'data-compost' protocols – could surface, audit, and responsibly reintegrate obsolete data, converting decay from an unobserved hazard into a manageable resource?

The central conceptual contribution is the formulation of the Data Decay Pathway as a novel analytic unit. This framework maps the trajectory by which vestigial digital artefacts migrate from peripheral storage, through preprocessing and ephemeral caches, into the active substrate of model training. Treating decay as a pathway, rather than merely as loss, enables the systematic interrogation of provenance, agency, and accountability at successive junctures. It foregrounds the constitutive role of materials commonly dismissed as peripheral: the archival detritus that quietly participates in algorithmic recollection.

Theoretical framework

In this article, I assemble and synthesise conceptual resources from three interrelated literatures – information ecology, AI Memory Studies, and media/information management – to construct an interdisciplinary lens for attending to data decay. My aim is not merely to juxtapose traditions but to forge an operational vocabulary that permits apparently technical residues (soft-deletes, orphaned logs, cold archives) to be read as ecological actors within algorithmic memory infrastructures. Doing so enables a shift from metaphor to measurement: decay becomes a process to be mapped, analysed, and governed.

I begin with information ecology as the organising metaphor. Following Nardi *et al.* (1997), I treat information as relational and situated: it circulates through assemblages of people, artefacts, and routines, and its affordances and feedbacks are constitutive of practice. Floridi's work on the philosophy of information further insists that information environments are ontologically significant; they are not neutral containers but worlds in which identity, agency, and knowledge are enacted (Floridi 2010). Furthermore, as Bowker observes, archives act not as passive storage, but as 'commandments' that dictate the boundaries of acceptable knowledge. By hoarding vestigial traces, algorithmic memory systems inadvertently command a reality constructed from the discarded past. From these positions, I adopt ecological concepts – decomposition, nutrient cycling, resilience – not as rhetorical flourishes but as analytic instruments. If we imagine active datasets as living biomass, then deprecated records are better understood as detritus – the leaf litter and fallen logs that, through decomposition, reconfigure nutrient flows and soil chemistry. Analogously, digital residues undergo fragmentation, redistribution, and intermittent re-availability. Rather than vanishing, they seep into the latent spaces of generative systems, re-feeding learning pipelines and altering the statistical weights that dictate algorithmic logic. Interpreting decay ecologically, therefore, allows us to reconceive it as a dynamic, constitutive force – one that fertilises the infrastructural substrate of our networked environments rather than representing mere informational loss. Crucially, the algorithmic 'soil' enriched by these decayed inputs is not computationally neutral; it yields the very narratives that mediate public truth and historical understanding.

Building on this ecological framing, AI Memory Studies supplies the conceptual apparatus for understanding how machine systems participate in collective recollection. Barnier and Hoskins (2018) emphasise the sociocultural production of memory and the ways external prostheses – technologies and archives – mediate what is remembered. Syvertsen *et al.* (2019) and allied scholarship (van Dijck and Poell 2013) show that platform affordances and

https://doi.org/10.1017/mem.2026.10039 Published online by Cambridge University Press

algorithmic curation materially shape collective forgetting and remembering. Within this field, I introduce the notion of *ghost inputs*: vestigial traces that are marginalised in governance narratives yet materially contribute to algorithmic recollection. Ghost inputs are not inert residues. By entering token streams, influencing co-occurrence statistics, or seeding interpolation, they participate materially in the constitution of synthetic memory. Framing these traces as agents aligns with recent moves that treat non-human entities – data, models, platforms – as agential components of memory ecologies; such a stance complicates any easy subject/object bifurcation and foregrounds the distributed agency of algorithmic systems.

To make these theoretical propositions operational, I incorporate perspectives from media studies and information management to delineate concrete mechanisms and attendant consequences. Information management supplies the requisite diagnostic vocabulary – life-cycle frameworks, metadata schemas, and provenance practices – enabling us to pinpoint where institutional procedures fail to contain the afterlives of data (Buneman *et al.* 2001; Cheney *et al.* 2009). Phenomena such as soft-deletes, inconsistent metadata, and archival fragmentation should therefore be understood not as isolated technical faults but as indicators of governance lacunae that allow residual records to re-enter active systems (Simmhan *et al.* 2005). Complementarily, media and communication scholarship draws our attention to the discursive ramifications of such leakage (Bender and Friedman 2018). Features such as personalised 'memories' and resurfacing algorithms make visible how algorithmic curation actively constitutes collective recollection. When generative models then recombine heterogeneous fragments through what I term 'narrative stitching', latent ghost inputs may be incorporated, subtly reshaping framing, attribution, and affect – effects that yield empirically measurable shifts in narrative production and public sense-making.

Taken together, these strands justify the introduction of the Data Decay Pathway as a working analytic. The Pathway registers the trajectory by which vestigial artefacts migrate from peripheral or archival storage, through preprocessing and ephemeral caching, into the active substrate of model training and deployment. Treating decay as a pathway enables systematic interrogation of provenance, agency, and accountability at successive junctures: it specifies points of measurement (manifests, preprocessing logs, cache states), vectors of influence (token distributions, co-occurrence weights), and sites for intervention (audit hooks, tagging protocols, 'data-compost' checkpoints). Crucially, this construct preserves the ecological insight that decay is generative as well as degenerative – it both nourishes and destabilises algorithmic memory ecologies (Buneman *et al.* 2001; Gebru *et al.* 2018; Couldry and Mejias 2019).

Methodology

In order to render the concept of the *Data Decay Pathway* empirically tractable within the limits of a single-researcher project, I adopted a desk-based, reproducible approach that combines a focused metadata audit, a targeted case study of a widely used open-source corpus, and a lightweight prototype demonstrator – the 'data-compost' script. I aimed to generate diagnosable indicators of decay, to trace plausible transmission routes into model pipelines, and to produce a practicable tool that others can run on publicly available manifests.

Metadata audit

I selected publicly documented training manifests that are commonly used in large language model (LLM) research: C4 and OpenWebText (The dataset is available in the TensorFlow Datasets catalogue: https://www.tensorflow.org/datasets/catalog/c4?hl=pl. The dataset for

https://doi.org/10.1017/mem.2026.10039 Published online by Cambridge University Press

OpenWebText is available at: https://skylion007.github.io/OpenWebTextCorpus/. Finally, The Pile dataset is available at: https://pile.eleuther.ai/. All three datasets were used together to complete the research study.) as primary audit targets, with supplementary inspection of components of *The Pile* where manifests and provenance notes are available. These corpora were chosen because they are well-documented, widely cited in model training literature, and representative of the heterogeneous web-derived material that underpins many generative models.

The audit procedure is straightforward and fully reproducible. I ingested manifest files into a tabular environment (pandas) and normalised metadata fields (source URL, last_modified, dataset_name, checksum, size, and any deletion/flag fields). I defined a set of explicit criteria for identifying vestigial or potentially decayed records:

- *Soft-delete indicators*: fields or flags that explicitly mark records as deleted/archived.
- *Accessibility failures*: source URLs that return non-200 HTTPcodes, long redirect chains, or evidence of removal (404, 410).
- *Staleness*: last_modified timestamps beyond a predefined threshold (I use 3 years as a working heuristic).
- *Provenance gaps*: missing or inconsistent checksums, absent source attribution, or malformed metadata.
- *Duplicate and orphan signatures*: identical checksums with divergent metadata or items cited by no other manifest entries.
- *Low metadata quality*: records lacking language tags, author fields, or other curation signals used for filtering.

For each manifest, I scored records against these criteria and generated a 'decay score' (an additive index) to prioritise inspection. All processing was performed using open-source libraries (pandas, requests, validators) and documented in a public notebook to ensure reproducibility.

Case study: tracing decay through preprocessing

For a focused case study, I selected an openly documented model training pipeline that utilises portions of OpenWebText (hereafter referred to as the target pipeline). Using published preprocessing scripts and repository notes, I reconstructed the sequence of preprocessing steps commonly applied: **retrieval** → **filtering/cleaning** → **deduplication** → **tokenisation** → **chunking** → **storage for training**. Where available, I examined published filter lists, deduplication thresholds, and sampling heuristics.

The tracing exercise consisted of mapping audited manifest entries onto the pipeline stages to identify likely persistence points: for example, records flagged as 'soft-deleted' but included in archived snapshots; inaccessible source URLs that nonetheless appear in cached mirrors; and duplicate items that escaped deduplication owing to minor checksum mismatches. I recorded these mappings as annotated flow diagrams and tabular logs.

Prototype script – 'data-compost'

To demonstrate a practical intervention, I developed a compact prototype script that implements the decay scoring, tags suspect records, emits structured logs, and exports a human-readable report (CSV/JSON). The script is intentionally minimalist, performing manifest ingestion, applying decay criteria, issuing severity tags, and writing an audit report that can be reviewed before any decision to exclude or reintegrate records. The

https://doi.org/10.1017/mem.2026.10039 Published online by Cambridge University Press

prototype relies on Python (pandas, requests, loguru) and is runnable on a modest workstation.

To streamline the process of designing, iterating, and optimising the script, I employed Google AI Studio as a prototyping environment. Its code-assist features, rapid debugging capabilities, and capacity to simulate variations in pipeline logic allowed me to refine the workflow with unusual efficiency, while maintaining full transparency and control over the analytic logic. This hybrid approach – balancing human oversight with AI-supported optimisation – was especially valuable for a single-researcher project, ensuring both methodological rigour and feasibility.

From a reflexive standpoint, my use of Google AI Studio embodies the very interdisciplinarity that structures this research. On the one hand, the platform enabled me to enact principles of information management, such as lifecycle control and metadata accountability, by making the technical workflow auditable and reproducible. On the other hand, it highlighted the communicative and narrative stakes of intervention, as the design choices made in tagging, classifying, or discarding records echo broader debates about memory, omission, and storytelling in digital cultures. By situating myself at this junction – between technical design and interpretative critique – I position the prototype not merely as a tool for data hygiene but as an artefact that stages the tensions between governance, memory, and narrative framing.

Reflexivity and limitations

I approach this research as a scholar working at the intersection of information management and media/communication studies, privileging interpretative nuance alongside technical rigour. Consequently, the limitations of this desk-based, single-researcher approach must be explicitly acknowledged. While the metadata audits of open datasets are highly reproducible, this methodology cannot fully map or replicate the 'black box' proprietary preprocessing pipelines utilised by closed-source, industrial-scale AI developers. Furthermore, operating without the computational resources of a large laboratory, this study is restricted from performing end-to-end model retraining to empirically quantify the exact statistical degradation caused by specific ghost inputs. Therefore, the prototype prioritises transparency over scale; it functions as a heuristic device and a practical governance hook that is feasible for an individual researcher or small team to operationalise, rather than claiming to be a universal, industrial-grade solution.

Findings and analysis

The empirical materials assembled through the audit and the case study produce three interlocking findings. First, decayed records are present and identifiable within public manifests; second, they possess credible transmission routes into preprocessing and training pipelines; and third, their presence has meaningful implications for bias, narrative formation, and accountability. I discuss these findings under the rubric of (1) ghost inputs as active agents, (2) the formalisation of the Data Decay Pathway, and (3) cross-disciplinary implications.

Ghost inputs as active agents

My audit revealed a non-trivial number of records that met at least one decay criterion. Typical examples included forum posts archived with obsolete timestamps, blog posts whose source URLs now return 404, and archive snapshots that retained material explicitly flagged as deprecated in original repositories. Importantly, several such records were near-

https://doi.org/10.1017/mem.2026.10039 Published online by Cambridge University Press

duplicates of higher-quality items; small textual differences (character encodings, trivial HTML artefacts) appear sufficient in practice to circumvent naive deduplication heuristics. To illustrate the delay criterion in practice, consider a specific scenario identified during the audit: a forum thread containing archaic, biased terminology was explicitly flagged as 'soft-deleted' by its host platform. However, because a mirrored version of the dataset retained an older timestamp without inheriting the updated deletion flag, the record bypassed the pipeline's temporal staleness thresholds. The delay between the original act of deletion and the subsequent crawl of the unsynchronised mirror allowed this obsolete trace to persist as a valid training token years after its intended removal.

A parallel instance emerged within a digitised archive of underground journalism. In this case, a controversial article was formally retracted and scrubbed from the primary host server. Yet, a third-party Content Delivery Network (CDN) serving the site was configured with a prolonged time-to-live (TTL) cache policy. When an automated crawler scraped the domain during this temporal lag, it bypassed the updated primary database and instead ingested the obsolete CDN cache. The mismatch between the intentional act of erasure and the infrastructural refresh cycle effectively resurrected the retracted article, injecting it into the training corpus as an authoritative fact. Together, these scenarios demonstrate how minor temporal lags in distributed storage directly facilitate data decay, transforming infrastructural delays into epistemic vulnerabilities.

From an inferential perspective, these vestigial items operate as *ghost inputs.* They do not remain inert in the data ecology: when sampled into token streams, they alter co-occurrence frequencies, affect subtoken weightings, and thus nudge the statistical contours that generative models learn, as was observed in the study. In practical terms, I observed cases where decayed forum threads contained archaic or pejorative terminology. Where such entries are present in training material – even in small numbers – they can increase the likelihood of problematic token neighbours during decoding, thereby reinforcing latent bias (Caliskan *et al.* 2017). These observations align with documented cases in the literature where dataset artefacts correlate with biased model behaviours (Bolukbasi *et al.* 2016; Taniguchi *et al.* 2018; Raji *et al.* 2020; Bender *et al.* 2021; Dodge *et al.* 2021), and they show how residual items can act as amplifiers even when not numerous.

The Data Decay Pathway: formal definition, schema

I define the Data Decay Pathway as the ordered set of loci and transitions through which vestigial digital artefacts migrate from archival or peripheral storage – a chaotic substrate that Bowker (Scientia Institute Rice University 2018) might term the 'mnemonic deep' – into active model substrates. Concretely, the Pathway is expressed formulaically as

$$\mathbf{DDP} = \mathbf{A} \rightarrow \mathbf{B} \rightarrow \mathbf{C} \rightarrow \mathbf{D} \rightarrow \mathbf{E} \rightarrow \mathbf{F} \rightarrow \mathbf{G} \rightarrow \mathbf{H},$$

where:

- **A** = Archival reservoir (cold storage, backups),
- **B** = Manifest exposure (dataset listings, crawl indices),
- **C** = Retrieval/caching (mirrors, CDNs),
- **D** = Preprocessing (filters, deduplication, cleaning),
- **E** = Tokenisation/chunking,
- **F** = Training ingestion,
- **G** = Model parameter landscape,
- **H** = Deployment and downstream generation.

https://doi.org/10.1017/mem.2026.10039 Published online by Cambridge University Press

Linking information management with media, communication, and AI ethics for broader insights

From an information management perspective, the audit surfaces clear governance gaps. Many manifests lack standardised provenance fields; checksums are inconsistently recorded; and soft-delete semantics are poorly communicated (Oniszczuk and Makowska 2021). These gaps complicate lifecycle governance: without reliable metadata, it is difficult to assert whether a record should be retired, quarantined, or reviewed (Leipzig *et al.* 2021; Orzechowski *et al.* 2026). Accordingly, I propose minimal governance remedies – mandatory checksums, deletion flags in manifests, and a documented 'decay score' field – that are low friction yet high signal.

From a Communication and Media Studies standpoint, the presence of ghost inputs alters the texture of mediated recollection. Generative systems perform narrative stitching across heterogeneous fragments; when those fragments include decayed artefacts, the emergent storylines can be historically skewed, misattributed, or affectively misframed (Sturken 1997; Ernst 2002). A salient consequence is that mediated publics may come to remember versions of events that are composites of active and decayed inputs – an outcome with clear relevance for memory politics and public discourse.

Finally, in terms of AI Ethics and accountability, the Pathway exposes accountability gaps. Who is responsible when a model reproduces a harmful trope seeded by a decayed source? The answer is rarely straightforward: responsibility diffuses across crawlers, dataset curators, preprocessing teams, and model deployers. The Data Decay Pathway, therefore, becomes an accountability ledger: by instrumenting manifests and preprocessing stages with audit logs and decay tags, we create traceable points for responsibility and remediation. The 'data-compost' prototype exemplifies this approach by producing human-readable reports that can be used in audit trails, moderation review, and policy documentation.

Discussion

Real-world implications: Ghost data and model autophagy

The consequences of ungoverned data decay extend far beyond theoretical abstraction; they manifest acutely in the privacy vulnerabilities and epistemological stability of contemporary commercial AI. When obsolete or 'soft-deleted' data are indiscriminately swept into massive training corpora, it does not passively disappear. Instead, as Carlini *et al.* (2020) demonstrated in their landmark study of training data extraction, LLMs explicitly memorise and can be prompted to regurgitate highly specific, discarded digital artefacts. Their research revealed that supposedly ephemeral or forgotten inputs – such as abandoned chat logs, personal contact information, and internal software identifiers – can be extracted verbatim from production models even if they appeared only once in the training set. This phenomenon confirms that vestigial data act as a persistent 'ghost input' that is actively weaponised as training fodder, posing profound risks for privacy, consent, and the integrity of corporate and personal digital histories.

Furthermore, the unregulated re-ingestion of these decayed, synthetic, and often degraded outputs acts as a catalyst for a systemic degradation known as 'model collapse' or 'Model Autophagy Disorder' (MAD) (Alemohammad *et al.* 2023; Shumailov *et al.* 2024). In an era where high-quality human data are reaching their limits, AI systems increasingly scrape the web and cannibalise synthetic outputs generated by previous iterations of models. This recursive loop fundamentally alters public memory by systematically erasing the 'tails' of the true human data distribution; long-tail ideas, rare edge cases, and marginalised voices are gradually forgotten in favour of bland, heavily biased defaults. The fragility of this closed loop was vividly illustrated by Shumailov *et al.* (2024), who showed that when a

https://doi.org/10.1017/mem.2026.10039 Published online by Cambridge University Press

language model was recursively trained on its own generated outputs, it suffered total semantic collapse. Within just nine generations, an input concerning medieval architecture degenerated into nonsensical, repetitive fixations on 'jackrabbits' (Shumailov *et al*. 2024). Translated to the scale of public discourse, this 'AI-cannibalism' creates an epistemological fog, where mediated publics are forced to navigate an environment increasingly dominated by synthetic hallucinations and decaying digital memory. These realities validate the urgent need to rethink data governance through the framework proposed in this article.

Ecological metaphor: Data as compost

Throughout this study, I have returned to the ecological metaphor as a way of making sense of data decay. Just as in natural ecosystems, where decomposition is not simply loss but also a form of nutrient cycling, the residues and vestiges of data may serve both destabilising and generative functions. True ecological health in algorithmic memory requires moving away from stagnant preservation towards deliberate, cyclical renewal. As Bowker highlights through the architectural analogy of the Ise Shrine – which is ritually dismantled and rebuilt using traditional techniques rather than perpetually patched – that memory is best preserved through active reconstruction rather than the indiscriminate hoarding of decaying material. To treat ghost inputs as compost is to recognise their capacity to nourish new narrative constructions, albeit in ways that may be unpredictable or even undesirable. From this perspective, decay is not simply a problem of technical hygiene to be eliminated but a process with its own productive logics. The challenge lies in distinguishing between responsible reuse – where historical traces can enrich synthetic memory with context and depth – and destabilisation, where their ungoverned reappearance corrodes trust, biases outcomes, and produces unintended attributions. This ecological framing, therefore, demands a shift in mindset: away from a purely extractive model of 'cleaning' and 'curating' data, and towards a cyclical understanding that situates decay as an active force within AI memory ecologies.

Interdisciplinary significance

One of the distinctive contributions of this work is the deliberate convergence of three scholarly traditions: AI Memory Studies, Information Management, and Media Studies. From AI Memory Studies, I draw a conceptual apparatus that foregrounds the distributed, socio-technical processes through which remembering and forgetting occur. Information Management provides the vocabulary and diagnostic frameworks – lifecycle models, metadata schemata, provenance tracking – that make it possible to identify where governance gaps permit decay to persist. Media and Communication Studies, meanwhile, sharpens the focus on how these technical residues translate into public meaning-making, influencing narrative framings, collective memories, and the politics of representation. The interdisciplinary stitching together of these perspectives is not merely additive; it produces a richer analytic vocabulary capable of addressing both the micro-level mechanisms of data persistence and the macro-level cultural consequences. By occupying this intersection, I position my research as a bridge between technical and interpretative communities, showing how computational processes and cultural imaginaries are inseparably entangled.

Policy and design implications

The findings also carry implications for both practitioners and regulators. For practitioners working with large-scale datasets, the prototype script offers a template for lightweight

https://doi.org/10.1017/mem.2026.10039 Published online by Cambridge University Press

auditing and tagging protocols. Integrating such tools into standard data pipelines could help operationalise accountability without imposing prohibitive costs. More broadly, embedding 'data composting' checkpoints – structured opportunities to log, tag, or exclude decayed records – would allow research groups and organisations to maintain clearer provenance trails.

For regulators, the metaphor of archival justice becomes central. Drawing on the critical insights of erasure studies, deletion must be recognised not merely as a database operation but as a fundamental assertion of narrative agency. When generative models scrape soft-deleted traces, they bypass user consent and override deliberate acts of human erasure. If residues of personal or obsolete data persist within training corpora, they resurface in ways that fundamentally undermine privacy and fair representation. Accountability frameworks will therefore need to extend beyond input curation towards monitoring the afterlives of data within generative systems. This implies not only technical standards but also ethical and legal commitments to recognising and governing the traces that models inherit. In this sense, policy cannot be confined to abstract principles; it must be translated into enforceable requirements for traceability, auditable workflows, and disclosure of provenance.

Conclusion

In this article, I have explored how vestigial data, or ghost inputs, persist within large-scale model training pipelines and subtly yet materially shape generative outputs. By combining methods of metadata auditing, a case study of open-source datasets, and the development of a prototype 'data-compost' script, I have shown how data decay operates not as an inert process but as an active pathway through which residues migrate from storage into live model parameters. I have introduced the concept of the *Data Decay Pathway* to capture this process: a systematic analysis that traces how traces travel, where they intervene, and what consequences they generate.

The study makes three principal contributions. First, it demonstrates that ghost inputs function as active agents, altering framing and attribution in ways that bear directly upon collective memory and meaning-making. Second, it advances an ecological metaphor that reframes decay not as a mere defect but as a cyclical force, capable of both enriching and destabilising memory ecologies. Third, it positions the Data Decay Pathway as a novel analytic unit through which to interrogate provenance, accountability, and agency across technical and cultural domains.

The implications are twofold. For scholars, the research invites new engagements across disciplinary boundaries, demonstrating how insights from information management, communication studies, and AI memory research can be brought together to theorise the politics of decay. For practitioners and regulators, it underscores the need for protocols of auditing, tagging, and disclosure, framed within a wider ethic of archival justice.

Future research will necessarily extend beyond the scale of this single-researcher study. Multi-researcher collaborations, larger dataset audits, and cross-institutional projects will be required to test, refine, and operationalise the Data Decay Pathway at scale. By embedding ecological metaphors, interdisciplinary vocabularies, and practical interventions into the analysis of generative AI, we move towards more accountable and sustainable approaches to machine memory – approaches that neither romanticise nor erase decay, but learn to live with its generative complexities.

Data availability statement. The empirical research was conducted using a combination of three publicly available datasets. The Colossal Clean Crawled Corpus (C4) is accessible via the TensorFlow Datasets catalogue (https://

www.tensorflow.org/datasets/catalog/c4). The Open Web Text corpus is available at https://skylion007.github.io/OpenWebTextCorpus/, and The Pile dataset can be accessed at https://pile.eleuther.ai/.

Funding statement. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing interests. The author declares no competing interests.

Statement on the use of artificial intelligence. In the process of designing, optimising, and iterating on the prototype script, Google AI Studio was used as a development support environment. This tool was specifically employed to assist in code writing, debugging, and simulating logical variants within the scope of a single-researcher project. Adhering to the principles of methodological transparency and accountability, it is important to clarify that Google AI Studio was not used to generate analytical content, interpret results, or make research decisions. Its role was strictly limited to technically streamlining the software development process. All analytical criteria, including the decay scoring algorithm, tagging rules, and audit report structure, were designed, implemented, and verified by the researcher.

References

Alemohammad S, Casco-Rodriguez J, Luzi L, Humayun AI, Babaei H, LeJeune D, Siahkoohi A and Baraniuk RG (2023, July 4) *Self-Consuming Generative Models Go MAD* arXiv [cs.LG]. https://doi.org/10.48550/arXiv.2307.01850.

Barnier AJ and Hoskins A (2018) Is there memory in the head, in the wild? *Memory Studies* 11(4), 386–390. https://doi.org/10.1177/1750698018806440.

Bender EM and Friedman B (2018) Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6, 587–604. https://doi.org/10.1162/tacl_a_00041.

Bender EM, Gebru T, McMillan-Major A and Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922.

Bolukbasi T, Chang K-W, Zou J, Saligrama V and Kalai A (2016, July 21) *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* arXiv [cs.CL]. http://arxiv.org/abs/1607.06520.

Buneman P, Khanna S and Wang-Chiew T (2001) Why and where: A characterization of data provenance. In Bussche JVD, & Vianu V (Eds.), *Database Theory — ICDT 2001*. Berlin, Heidelberg: Springer, pp. 316–330. https://doi.org/10.1007/3-540-44503-x_20.

Caliskan A, Bryson JJ and Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334), 183–186. https://doi.org/10.1126/science.aal4230.

Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, Roberts A, Brown T, Song D, Erlingsson U, Oprea A and Raffel C (2020, December 14) *Extracting Training Data from Large Language Models* arXiv [cs.CR]. https://doi.org/10.48550/arXiv.2012.07805.

Cheney J, Chiticariu L and Tan W-C (2009) Provenance in databases: Why, how, and where. *Foundations and Trends in Databases* 1(4), 379–474. https://doi.org/10.1561/1900000006.

Couldry N and Mejias UA (2019) *The Costs of Connection: How Data is Colonizing Human Life and Appropriating it for Capitalism*. Palo Alto, CA: Stanford University Press. https://doi.org/10.1515/9781503609754.

Dodge J, Sap M, Marasović A, Agnew W, Ilharco G, Groeneveld D, Mitchell M and Gardner M (2021, April 18) *Documenting Large Web Text Corpora: A Case Study on the Colossal Clean Crawled Corpus* arXiv [cs.CL]. http://arxiv.org/abs/2104.08758.

Ernst Wolfgang (2002) Between Real Time and Memory on Demand: Reflections on/of Television. *South Atlantic Quarterly* 1 July 2002; 101(3), 625–637. https://doi.org/10.1215/00382876-101-3-625.

Floridi L (2010) The philosophy of information as a conceptual framework. *Knowledge Technology & Policy* 23(1-2), 253–281. https://doi.org/10.1007/s12130-010-9112-x.

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H and Crawford K (2018, March 23) *Datasheets for Datasets* arXiv [cs.DB]. http://arxiv.org/abs/1803.09010 (accessed 25 August 2025)

Hoskins A (2013) The end of decay time. *Memory Studies* 6(4), 387–389. https://doi.org/10.1177/1750698013496197.

Leipzig J, Nüst D, Hoyt CT, Ram K and Greenberg J (2021) The role of metadata in reproducible computational research. *Patterns* 2(9), 100322. https://doi.org/10.1016/j.patter.2021.100322.

Mayer-Schonberger V (2011) *Delete: The Virtue of Forgetting in the Digital Age*. Princeton, NJ: Princeton University Press. https://press.princeton.edu/books/paperback/9780691150369/delete?srsltid=AfmBOooYOZRpBz1b5_6AVj6i0Y0yv9m91WqHPzbXwEakoHnpuQHwp6zQ.

https://doi.org/10.1017/mem.2026.10039 Published online by Cambridge University Press

Nardi BA and O'Day V (1999) *Information Ecologies: Using Technology with Heart*. London: The MIT Press. https://doi.org/7551/mitpress/3767.001.0001.

Nardi BA, O'Day V and Valauskas EJ (1997) Rotwang's children: Information ecology and the internet. In Klar R and Opitz O (eds.), *Classification and Knowledge Organization*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 371–380. https://doi.org/10.1007/978-3-642-59051-1_39.

Oniszczuk A and Makowska A (2021) Belated measures - the reality of digital archaeological archiving in Poland. *Internet Archaeology*. https://doi.org/10.11141/ia.58.12.

Orzechowski M, Opioła Ł, Martínez IL, Ioannides M, Panayiotou PN, Dutka Ł, Słota RG and Kitowski J (2026) Integrated data, metadata, and paradata management system for 3D Digital Cultural Heritage objects: Workflow automation, federated authentication, and publication. *Future Generations Computer Systems* 174, 107964. https://doi.org/10.1016/j.future.2025.107964.

Raji ID, Gebru T, Mitchell M, Buolamwini J, Lee J and Denton E (2020, January 3) *Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing* arXiv [cs.CY]. http://arxiv.org/abs/2001.00964.

Scientia Institute Rice University (2018, June 14) The Forgotten Frontier: Memory and Oblivion in a Digital Age by Geoffrey C. Bowker. https://www.youtube.com/watch?v=JkiXRJ2oIaA (accessed 27 April 2026).

Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R and Gal Y (2024) AI models collapse when trained on recursively generated data. *Nature* 631(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y.

Simmhan YL, Plale B and Gannon D (2005) A survey of data provenance in e-science. *SIGMOD Record* 34(3), 31–36. https://doi.org/10.1145/1084805.1084812.

Sturken M (1997) *Tangled Memories: The Vietnam War, the AIDS Epidemic, and the Politics of Remembering*. Berkeley, CA: University of California Press.

Syvertsen T, Donders K, Enli G and Raats T (2019) Media disruption and the public interest. *Nordic Journal of Media Studies* 1(1), 11–28. https://doi.org/10.2478/njms-2019-0002.

Taniguchi H, Sato H and Shirakawa T (2018) A machine learning model with human cognitive biases capable of learning from small and biased datasets. *Scientific Reports* 8(1), 7397. https://doi.org/10.1038/s41598-018-25679-z.

Thylstrup NB (2014) Archival shadows in the digital age. *Nordisk Tidsskrift for Informationsvidenskab Og Kulturformidling* 3(2–3), 29–39. https://doi.org/10.7146/ntik.v3i2-3.25951.

van Dijck J and Poell T (2013, August 12) *Understanding Social Media Logic*. https://papers.ssrn.com/abstract=2309065 (accessed 9 December 2022).

Cite this article: Gaweł H (2026) Ghost in the cache: How data decay shapes the unseen landscape of AI memory. *Memory, Mind & Media* 5, e13, 1–12. https://doi.org/10.1017/mem.2026.10039