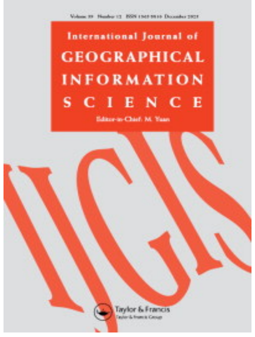

Taylor & Francis
Taylor & Francis Group

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: www.tandfonline.com/journals/tgis20

GeoAI collapse? Ethical implications of synthetic geospatial data use

Bo Zhao & Yue Lin

To cite this article: Bo Zhao & Yue Lin (05 Jan 2026): GeoAI collapse? Ethical implications of synthetic geospatial data use, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2025.2609806

To link to this article: https://doi.org/10.1080/13658816.2025.2609806

Published online: 05 Jan 2026.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2025.2609806

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

GeoAI collapse? Ethical implications of synthetic geospatial data use

Bo Zhao[a] and Yue Lin[b]

[a]Department of Geography, University of Washington, Seattle, WA, USA; [b]Department of Geography and Geographic Information Science, University of Illinois Urbana-Champaign, Urbana, IL, USA

ABSTRACT

As synthetic data become increasingly embedded in GeoAI workflows, the long-term risks of recursive training on AI-generated inputs remain insufficiently understood. This study introduces and empirically examines the concept of *GeoAI collapse* – a degenerative process in which repeated reliance on synthetic geospatial data leads to progressive performance degradation. Focusing on semantic segmentation of street-level imagery, we simulated multigenerational training cycles using a conditional generative adversarial network *pix2pix*. Results demonstrated substantial declines in both visual fidelity and quantitative performance metrics, with rare place-based features exhibiting near-total collapse by later generations. These findings reveal structural vulnerabilities in GeoAI pipelines and highlight the ethical risks posed by unscreened, synthetic data entering public geospatial datasets. Beyond technical degradation, the study situates these risks within a broader phenomenon of digital placelessness – the erosion of geographic specificity and contextual meaning as AI-generated representations progressively abstract the lived realities of place. To address these challenges, we propose provenance-aware evaluation protocols that emphasize resilience, spatial fairness, and transparency. This work calls for a critical reframing of synthetic data practices in GeoAI to safeguard the integrity of geographic knowledge production.

ARTICLE HISTORY
Received 24 June 2025
Accepted 22 December 2025

KEYWORDS
GeoAI; ethics; model collapse

Introduction

Today's geospatial technologies are undergoing a profound transformation – what some have begun to characterize as a posthuman shift (Lin and Zhao 2025). In such a shift, the production of geographic knowledge is increasingly co-shaped by human judgment and artificial intelligence. Spatial representation and analysis are no longer only mediated by algorithms but are collaboratively produced by human and algorithmic agents. GeoAI – geospatial artificial intelligence – has emerged at the heart of this moment, promising new modes of automated spatial reasoning and data-driven decision-making. Yet as GeoAI gains momentum, there is a growing concern that its rapid

CONTACT Bo Zhao zhaobo@uw.edu

expansion has been accompanied by insufficient scrutiny of its ethical foundations (Li *et al*. 2024, Marasinghe *et al*. 2024, Oluoch 2024). This paper specifically investigates one emerging source of concern: the ethical implications of the growing reliance on synthetic geospatial data, defined here as data generated not through direct observation, but through machine (in particular, AI) synthesis. In contexts where labeled training data are scarce, synthetic geospatial datasets allow GIScientists and geographers to augment their models with statistically plausible yet artificial samples. These data are especially appealing when privacy constraints limit the use of real-world data, or when acquiring such data are prohibitively expensive or labor-intensive (Marwala *et al*. 2023). Moreover, synthetic data support the optimization of AI efficiency and training throughput – considerations that have become increasingly salient in the face of rising computational and environmental costs (Li *et al* 2025). Taken together, these factors have led to a growing number of GeoAI pipelines that now depend – at least in part – on data that are machine-generated.

However, these practical benefits carry a profound risk. When synthetic data are used recursively – training new models on outputs of previous ones – the AI ecosystem may become increasingly self-referential. This recursive feedback loop poses a structural threat to model robustness and diversity. Recent studies (Shumailov *et al*. 2024, Burden *et al*. 2024) show that such a loop gradually amplifies dominant statistical patterns while filtering out rare or complex features, narrowing the diversity of learned representations. Over successive cycles, models begin to produce their own artifacts rather than the true distribution of the world, leading to homogenized outputs and epistemic drift. This phenomenon is known as 'model collapse', where models trained on synthetic outputs of other models begin to exhibit degraded performance, lose generalization capacity, and reproduce narrow, homogenized outputs. In the context of GeoAI, while most current applications still rely primarily on real-world data, the rapid growth of synthetic geospatial imagery in open repositories raises a forward-looking concern: over time, such data could be re-used – sometimes unknowingly – for model retraining, creating the recursive exposure we simulate in this study (Walker and Winders 2025). Left unchecked, these vulnerabilities could propagate silently through GIS-related applications.

Beyond these ethical concerns, it is important to emphasize that this paper does not argue against the use of synthetic data *per se*. In many cases – especially where data scarcity or privacy constraints exist – synthetic geospatial data can provide meaningful support for model development and research innovation. However, we contend that their use must be context-sensitive, auditable, and constrained by clear ethical safeguards. The uncritical use of synthetic data as interchangeable with real-world observations risks entrenching a form of geospatial simulation untethered from empirical reality.

In this study, we define 'GeoAI collapse' as the progressive degradation of model performance resulting from the recursive use of synthetic geospatial data in training GeoAI systems. We particularly focus on AI-generated imagery embodying geographical scenes in this study and hypothesize that such collapse manifests in two primary forms: (1) a general decline in semantic segmentation accuracy across generations, and (2) a disproportionate erosion in the model's ability to detect rare or lowfrequency place-based features. This paper empirically investigates these patterns and their ethical implications for the development of responsible GeoAI. It should be mentioned that the concept of GeoAI collapse is intentionally framed to be more broadly than one mechanism of recursive model degradation to reflect a systemic and evolving risk in the life cycle of GeoAI – where synthetic geospatial data, once introduced into open or shared datasets, may gradually shape model behavior, diversity, and epistemic grounding over time.

In the sections that follow, we first review current debates on the use of synthetic data in GeoAI, with attention to both technical drivers and ethical concerns. Building on this, we present a case study designed not to document an existing instance of GeoAI collapse, but to simulate a plausible and increasingly observable scenario in which recursive reliance on synthetic geospatial data could lead to systemic degradation. This experiment serves to illustrate how such a process might unfold if left unchecked, highlighting both the technical risks and the ethical challenges associated with synthetic data use. Rather than claiming that all GeoAI systems already face this problem, our goal is to encourage proactive reflection and accountability in how the community approaches synthetic data in the evolving landscape of GeoAI.

The relevant works

The existing synthetic data in GeoAI

Synthetic data – broadly defined as model-generated outputs used in place of or alongside 'real' empirically observed data – has long played a role in GIScience and related spatial modeling traditions. Historically, geographic researchers have relied on simulation-based methods and synthetic constructs to represent processes that could not be directly observed – from synthetic aperture radar and spatial interpolation to agent-based urban modeling (Goodchild 1992, Foody 2002). Later, advances in computation enabled more sophisticated forms of synthetic data generation, such as the use of social media and mobility data to model urban dynamics (Wu *et al.* 2018) and the development of digital twins and simulation frameworks to explore synthetic urban systems (Papyshev and Yarime 2021). In addition, researchers have also adopted traditional data-augmentation techniques – such as rotation, scaling, or projection transformations (Hu *et al.* 2022) – to generate additional training samples from existing observations and enhance model robustness. With the emergence of GeoAI, however, these established synthetic practices have been radically transformed by generative algorithms that produce data not merely through abstraction or interpolation, but through autonomous synthesis (Li *et al.* 2024).

These algorithmic approaches addressed situations where fine-grained real data were scarce or inaccessible due to privacy constraints. In this paper, however, we focus on a narrower conception of synthetic data: AI-generated data used to train AI models (Jacobsen 2023). Over the past decade, the rise of deep learning and GeoAI has introduced new methods to produce such data. Geospatial models built on convolutional neural networks can automatically detect, classify, or generate spatial features (e.g., building footprints) from remote sensing inputs, with the outputs often fed back into mapping platforms and databases (Heris *et al.* 2020, Guo *et al.* 2021, Lin *et al.* 2023).

The advent of modern generative models has further accelerated the production of synthetic geospatial data at scale. Techniques such as variational autoencoders and generative adversarial networks (GANs) have been applied to create realistic map images, street scenes, and multispectral satellite imagery (Kim and Bansal 2023, Alibani *et al*. 2024). More recently, foundation models like GPT-5, Gemini, Llama, and DALL·E have enabled on-demand generation of text, images, and other geospatially relevant content, normalizing synthetic data as a routine part of GeoAI workflows (Romano 2025). Today, many sources of geographic information – from social media and user-contributed reviews to wikis and even academic publications – are increasingly populated with AI-generated content (Knight *et al*. 2023, Brooks *et al*. 2024, Kendall and Teixeira da Silva 2024, Wei and Tyson 2024). This ubiquity of synthetic data reflects a new culture of data creation – one increasingly mediated by algorithms and detached from direct human observation (Crawford 2021). Within this emerging landscape, artificially generated information is routinely intermingled with human-origin data across the web and readily scraped into GeoAI training datasets.

The turn to synthetic data in GeoAI has been driven by several practical and ethical motivations. Foremost is the promise that synthetic datasets can provide rich variability and coverage of rare events that real-world data collection might miss (Nikolenko 2021). By generating diverse hypothetical samples (for instance, simulating uncommon disaster scenarios or rare landscape patterns), researchers can expose models to a broader spectrum of cases than would otherwise be inaccessible through empirical observation alone. Incorporating carefully generated synthetic samples alongside real data, as shown in existing studies, can improve model robustness and generalization by increasing input diversity (Jacobsen 2023, Xu *et al*. 2024). Privacy and confidentiality concern also motivate synthetic data creation: in domains like human mobility or demographics, artificially generated or geomasked (i.e., spatially altered to obscure exact locations) populations and human mobility traces stand in for sensitive real personal data, protecting individual privacy while enabling analysis (Wu *et al*. 2018, Lin 2023, Prédhumeau and Manley 2023). A related benefit is overcoming data scarcity and access limitations – for example, when only limited ground-truth labels or observations exist for training, synthetic augmentation can fill the gaps (Zhu *et al*. 2023, Amiri *et al*. 2024). In addition, synthetic data offer advantages in scalability and cost: once a generative procedure is set up, virtually unlimited training data can be produced without the expense of field collection or manual labeling (Li *et al*. 2021). This is particularly valuable for geospatial deep learning tasks that demand massive, labeled datasets (e.g., road detection or place name recognition).

In short, synthetic geospatial data are seen to bolster model performance while reducing traditional data collection constraints – a win-win that explains their growing appeal in GeoAI research and applications. Despite these advancements, little attention has been paid to the long-term impacts of recursive synthetic data use within GeoAI systems themselves. While prior studies have explored the technical benefits of synthetic data augmentation and its privacy-preserving features (Hu *et al*. 2022, Lin and Xiao 2023), few have empirically tested whether recursive generation degrades performance in geospatial contexts, or how such degradation might manifest in tasks requiring spatial fidelity and semantic precision.

Model collapse effects and its relevance to GeoAI

Recent studies in AI have identified a phenomenon known as 'model collapse', a degenerative process in which model performance and output quality degrade over iterative cycles of training on synthetic data (Shumailov *et al*. 2024). Essentially, when a model 'learns' predominantly from the by-products of other models rather than from independent observations, it can enter a feedback loop that amplifies errors and reduces data diversity. In computer vision, for example, Hataya *et al*. (2023) demonstrated that when generative models were repeatedly trained on their own output (e.g. images produced by earlier model versions), the resulting images became increasingly distorted and homogeneous, and downstream tasks like image classification suffer noticeable performance drops. Other vision researchers report similar findings: the variability of generated images diminishes with each generation of training, yielding an eventual loss of fidelity akin to repeatedly photocopying an image (Yamaguchi and Fukuda 2023, Xing *et al*. 2025).

This same risk has been observed in natural language processing. Guo *et al*. (2024) showed that large language models trained on synthetic text began to exhibit a decline in linguistic diversity, losing the richness of expression found in human text over successive iterations. In practical terms, a language model that consumes too much AI-written text may start outputting more formulaic, less accurate responses – an effect noted by Shumailov *et al*. (2024) and echoed in industry discussions of 'LLM degeneration'. Indeed, one Nature news article described how an LLM quickly started to 'spew nonsense' after being trained on AI-generated outputs, underscoring how fast collapse can set in Gibney (2024). Beyond perceptible nonsense, there is a deeper concern that recursively trained models cease to generalize correctly because their training data no longer reflects the true complexity of the world (Wenger 2024). The issue extends to knowledge bases and information retrieval as well. If a significant fraction of online geographic knowledge is AI-generated (and potentially error-laden), an AI model that later trains on that knowledge might internalize those errors, leading to what Peterson (2025) termed 'knowledge collapse'. In short, the unchecked use of synthetic data in training can induce a self-perpetuating decline in both the diversity and accuracy of AI outputs. This emergent risk was once hypothetical but is now supported by empirical evidence across vision and language domains.

We argue that GeoAI is especially vulnerable to model collapse and related risks, due to a confluence of factors inherent to GeoAI. First, contemporary GeoAI methods are heavily reliant on pretrained foundation models and web-scale datasets that lie outside the direct control of the geospatial community (Li and Ning 2023, Manvi *et al*. 2024). Many GeoAI tools – from map assistants (Zhang *et al*. 2024a) to GIS chatbots (Zhang *et al*. 2024b) – are built by fine-tuning general LLMs or vision models that have already been exposed to large amounts of synthetic content. This means GeoAI applications often start from a foundation that may be partly compromised by earlier rounds of AI-generated data. If those base models have begun to 'collapse' in their respective domains (e.g. an LLM with subtly degraded understanding of factual geographic knowledge), the geospatial application will inherit those weaknesses.

Second, GeoAI's dependence on open, crowd-sourced, and opportunistic data makes it hard to insulate its training pipeline from synthetic data infiltration.

Geospatial model training often draws on sources like OpenStreetMap (OSM), street-level images, social media, news, and academic literature to obtain labeled examples or ancillary features. Each of these sources is now intermixed with AI-generated information (Brooks *et al*. 2024, Kendall and Teixeira da Silva 2024, Wei and Tyson 2024). For example, a recent analysis estimated that about 3.5% of building footprint data in OSM were AI-generated contributions as of 2024 (Fila *et al*. 2025). In contrast to a more controlled domain (say, medical imaging, where a curated set of hospital scans might be used), geospatial datasets tend to aggregate whatever data can be found for a region or topic. This wide net makes it practically infeasible to filter out synthetic entries, especially when they are not clearly marked. It is therefore essential to develop clear labeling conventions that indicate not only whether a contribution is AI-generated but also whether it was created in an AI-assisted, human-edited manner. Such distinction would help maintain interpretability and trust in the increasingly hybrid data environments.

Third, the geographic specificity of many GeoAI tasks heightens the potential damage from synthetic-data-induced errors, distortions or biases. Geospatial models must often capture fine-grained, local variations – for example, recognizing an outlier pattern in land use, or detecting an unusual trajectory in traffic data. If synthetic training data gloss over or systematically underrepresent outliers and minorities, the resulting models could lose the ability to detect these critical phenomena. Recent work provides evidence of this danger: Z. Li *et al*. (2025) found that when AI-generated labels were used for crop types in spatial datasets, the model's performance in identifying anomalous patterns significantly worsened, likely due to noise and uncertainty in those synthetic labels. Likewise, Romano (2025) discussed how generative spatial point datasets tended to be overly smooth and could miss extreme values or rare spatial outliers, which are often the very signals analysts care about for anomaly detection or hazard identification.

In spatial analysis, seemingly small distortions – a few meters error in location here, a subtle mode collapse in imagery there – can compound into large errors in spatial reasoning (e.g., misidentifying the site of a flood because the generative model never saw a levee breach scenario in training). Even the integrity of foundational geospatial databases might be at stake: for instance, if AI-generated building footprints in OSM have systematic shape errors or missing attributes, any model trained on OSM for urban analytics would propagate those errors (Fila *et al*. 2025). As shown, GeoAI's entwined relationship with broader AI ecosystems and its reliance on high-fidelity, spatially contextualized information render it uniquely susceptible to the risks posed by the unchecked use of synthetic geospatial datasets. Yet to our knowledge, these vulnerabilities have not been formally tested through controlled experiments. In the next section, we address this gap by presenting a case study that empirically evaluates the cumulative effects of recursive synthetic training on GeoAI model performance.

Data and methodologies

To empirically examine the potential collapse effects introduced by synthetic data in GeoAI, we designed a multi-generation image generation experiment using semantic-to-realistic conditional image translation models (generate images conditioned on semantic labels). We used AI-generated imagery as an illustration of synthetic data use in GeoAI. It is important to note, however, that synthetic geospatial data can take multiple forms. Model-derived analytical outputs (such as AI-extracted building footprints) and AI-generated images differ in how they are produced, yet both exemplify the broader category of artificially created geographic information whose reuse may propagate distortions or inherited bias through model training pipelines. In our designed experiment, the generated images depict detailed geographic scenes, reflecting a broader shift within GeoAI from abstract representations of *space* toward the recognition and interpretation of *place* – a direction deeply rooted in geography's longstanding tradition of place-based inquiry (Zhao 2022). As GeoAI increasingly centers on identifying and classifying geographic entities in specific contexts, this emphasis on *place* has become foundational to many of its applications.

The goal of our experiment was to simulate a plausible feedback loop scenario in which synthetic data – generated by earlier GeoAI models – are progressively reused to train subsequent generations, thereby testing for performance degradation over iterations. Importantly, this multi-generation framework was designed to emulate a real-world risk: the dissemination of unscreened synthetic data across the Internet or public repositories, where such content could be unknowingly incorporated into future training sets. This risk is particularly relevant to our experiment with street-level imagery, as nowadays major datasets are crowd-sourced and vulnerable to synthetic data from contributors. Our design thus reflects an increasingly likely condition in which synthetic data silently influence both model performance and data quality over time.

Our training dataset was prepared based on a combination of two well-established, high-quality real-world street-level imagery datasets: Mapillary Vistas (Neuhold *et al*. 2017) and Cityscapes (Cordts *et al*. 2016), which respectively provide finely annotated urban scenes from European cities and diverse street-level imagery from around the world. To support training across five-generation models in this experiment, we divided the full dataset into five equal subsets. Each subset contained 4195 real urban street-level images. Each image was paired with its corresponding semantic segmentation labels (i.e., the object class assigned to each pixel in the image), providing the structured input necessary for conditional image generation tasks. The training data cover a broad range of urban features such as roads, buildings, vegetation, sidewalks, vehicles, and pedestrians. The extensive coverage and rich annotation also make them ideal for establishing a baseline model trained exclusively on human-labeled data. The data were preprocessed to maintain consistency in resolution and class balance, ensuring that rare urban features (e.g., fences, riders, traffic lights) were preserved during training.

We used *pix2pix* (Isola *et al*. 2017), a conditional GAN (cGAN) widely adopted in recent GeoAI applications for urban analysis (Ito *et al*. 2024), to translate semantic segmentation labels into realistic-looking street-level images. The first-generation model (Gen-1) was trained entirely on 4,195 real image-label pairs from the first subset of the training data. To simulate recursive synthetic training, for each subsequent generation (Gen-*t*, where $t > 1$), the real images in the *t*-th subset were replaced with synthetic images generated by the previous generation model (Gen-*t*–*1*), conditioned on the corresponding semantic labels in the *t*-th subset. This recursive generation process was repeated until Gen-5. Each model was trained using a batch size of 8 over 200 epochs. The learning rate was set to 0.0002 for the first 100 epochs, followed by a linear decay to zero over the remaining 100 epochs to promote stable convergence and prevent overfitting in later stages of training. In each generation, the model was trained on the base of the previous generation to mimic real-world practices in which pretrained models are iteratively updated with newly acquired data. This design allows for direct observation of the cumulative effects of synthetic data over time and the degradation patterns attributable to repeated synthetic data reuse.

Quantitative evaluation of generative models like *pix2pix* is known to be challenging, but it has become a common practice in recent work to use pre-trained semantic classifiers to measure the fidelity of the generated images as a proxy metric (Isola *et al*. 2017). The rationale is that if the generated images are sufficiently realistic, classifiers trained on real data should still be able to identify objects in these images correctly. To this end, we used each of the five-generation models to generate street-level images from the semantic segmentation labels of 500 test samples. We then adopted OneFormer (Jain *et al*. 2023), a popular transformer-based image segmentation model, pretrained on real Mapillary Vistas and Cityscapes data, as a semantic classifier to evaluate the fidelity of generated images. Model performance was indicated using standard semantic segmentation metrics, specifically F1 score and Intersection over Union (IoU), computed per class. We ranked object classes by the number of pixels in the ground truth labels to enable clear analysis of how common versus rare place-based features were affected across generations.

Results

The results reveal a clear trajectory of performance degradation in both visual quality and quantitative metrics.This degradation is particularly pronounced in the detection and segmentation of rare place-based features.

Figure 1 presents a visual comparison across generations. Each row represents a different place in urban scene, progressing from the ground-truth semantic segmentation labels (far left), through the real-world street-level images, followed by synthetic outputs from Gen-1, Gen-3, and Gen-5. The degradation is clearly visible across scenes. In the first row, architectural detail is initially preserved but begins to blur and distort by Gen-3; by Gen-5, rooflines and windows dissolve into a foggy mass, and depth cues vanish. In the second row, what appears to be a tower or obelisk becomes increasingly difficult to discern, and the cityscape gradually transforms into an abstract, painterly haze. The third row offers a striking example of spatial degradation: straight tree-lined streets become warped and smoky, with tree shapes resembling plumes or brush strokes in later generations. In the fourth row, which originally includes clear sidewalk edges and multiple pedestrians, the human figures progressively dissolve. By Gen-5, the image has lost virtually all semantic detail, recalling the aesthetic of 19th-century impressionist paintings more than a street scene. These visual patterns confirm that recursive generation causes severe fidelity loss and semantic erosion.

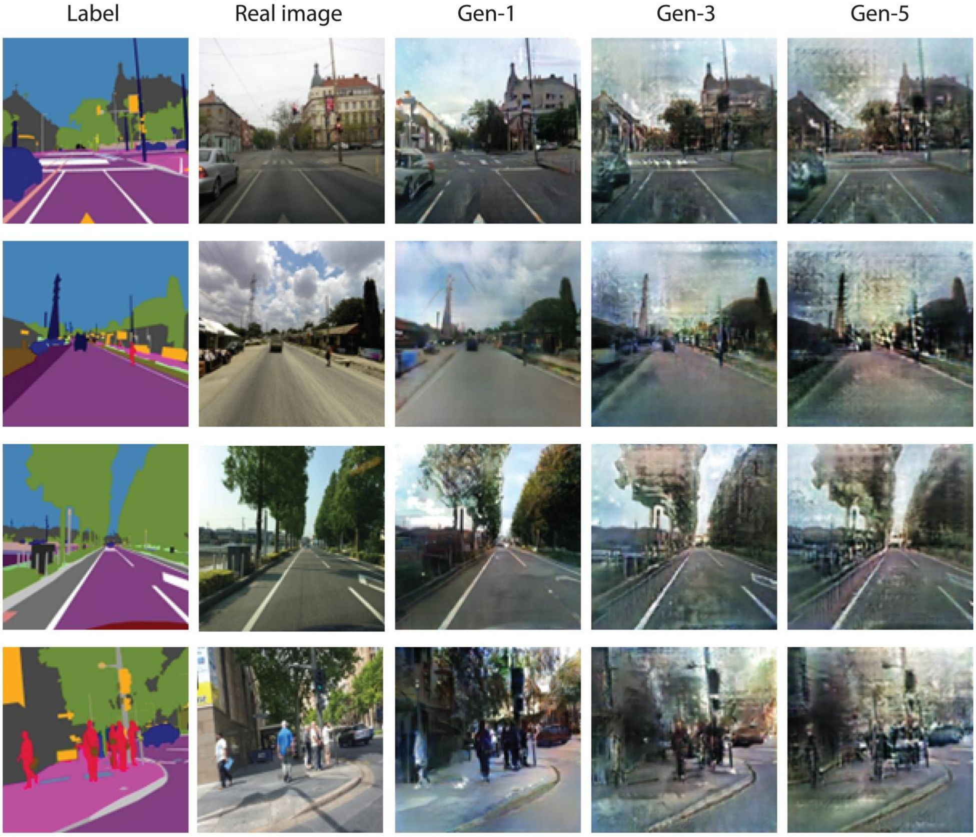

Figure 1. Visual comparison of street-level images over generations.

Quantitative results further substantiate the visual evidence. Figure 2 reports the F1 scores across a variety of place-based features over five generations of training. As expected, performance was highest when models were trained on real images. The outputs of Gen-1, which was trained on real data, showed fidelity comparable to the baseline real street-level images, as reflected in their generally closely aligned F1 scores. However, as models were trained on synthetic outputs from previous generations, the scores declined remarkably. The most drastic declines were observed in rare place-based features such as ‘rider’, ‘motorcycle’, and ‘traffic sign’, all of which dropped from baseline F1 scores of around 0.7 to below 0.1 by Gen-4 or Gen-5. Mid-frequency features such as ‘fence’, ‘sidewalk’, and ‘pole’ showed more moderate but still significant reductions. Common features like ‘road’ and ‘building’ retained relatively higher F1 values, especially in early generations, but nonetheless trended downward by Gen-4 and Gen-5. This suggests that recursive training undermines the model’s ability to generalize across class distributions, disproportionately affecting less frequently appeared place-based features.

Figure 3 shows a similar trend for IoU, which emphasizes spatial alignment and segmentation accuracy. The performance decay here was even more striking for boundary-sensitive categories. While place-based features like ‘car’ and ‘bus’ maintained some coherence in early generations, their IoU values sharply fell by Gen-4,

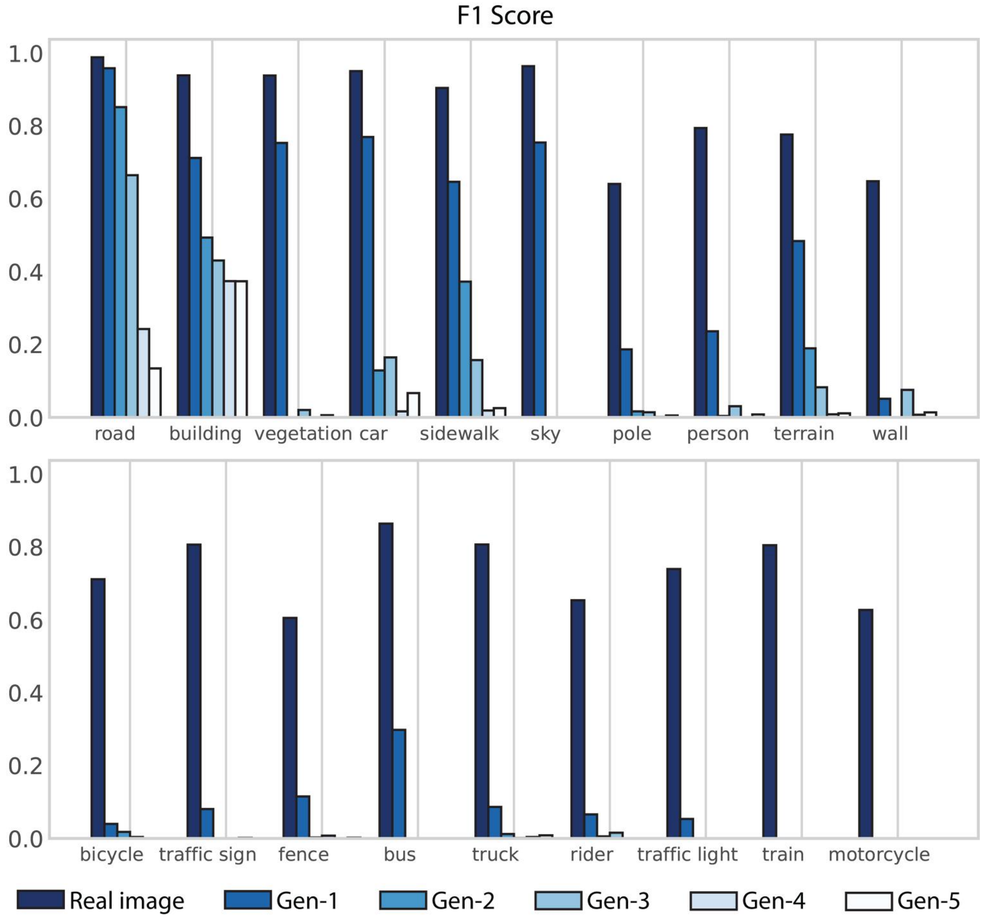

Figure 2. F1 scores by class over generations.

indicating blurred segmentation masks and object overlap. Rare and spatially constrained categories such as ‘rider’, ‘traffic light’, and ‘train’ exhibited near-total collapse, with IoU values approaching zero. These results not only reinforce the degradation observed in F1 scores but also highlight the model’s growing inability to localize and delineate object boundaries as it consumes recursively generated data. The results provide robust empirical evidence for the hypothesis of GeoAI collapse under recursive synthetic training. Visual artifacts accumulate visibly across generations, while F1 and IoU metrics demonstrate systematic and disproportionate erosion of model accuracy – particularly in rare, detailed, or small place-based features.

Discussions

Consequences of unscreened synthetic geospatial data

It is important to note that GeoAI collapse does not imply that all GeoAI researchers use synthetic data indiscriminately. In practice, responsible GIScientists routinely perform exploratory data analysis and quality checks before model training. Rather, our

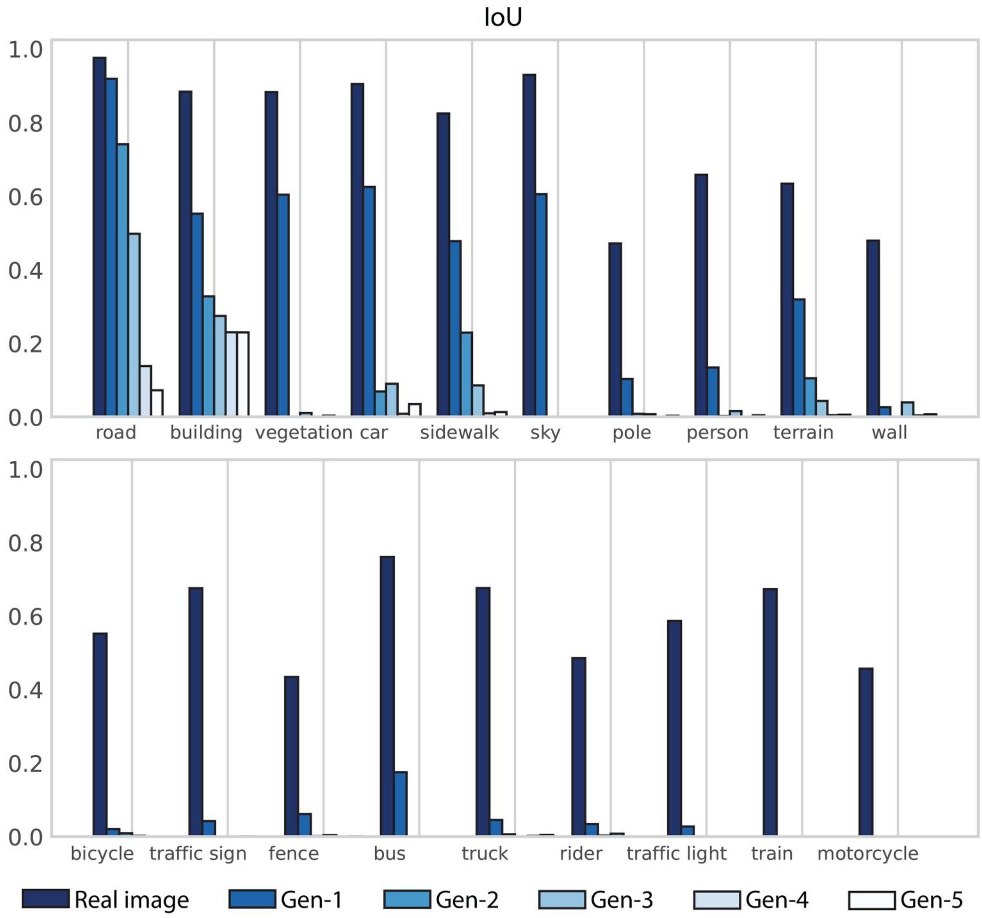

Figure 3. IoU by class over generations.

experiment simulates a plausible and increasingly observable scenario – one that is increasingly likely in the evolving landscape of GeoAI – in which synthetic geospatial data (e.g., street-level imagery) become embedded within public datasets (e.g., crowd-sourced mapping platforms) and are later reused for training by users unaware of its artificial origin. The GIScience community increasingly relies on open, collaborative data ecosystems – OSM, Mapillary, Wikipedia, and others – an evolution of what Goodchild (2007) termed 'volunteered geographic information (VGI)' – where the traditional expectation is that content reflects some form of ground truth (e.g., human observation or verifiable records). If these shared data platforms become unintentionally saturated with AI-generated contributions, the downstream effects on GeoAI models could be significant and difficult to detect. These findings should be understood not as evidence that all kinds of existing GeoAI systems already experience such recursive collapse, but as a scenario-based warning of a plausible risk. This scenario encapsulates the core mechanism of GeoAI collapse: as synthetic data circulate through open repositories and recursively feed new models, their distortions and biases can silently propagate, eroding both model diversity and the reliability of geographic knowledge over time.

There is emerging evidence that this risk is not merely hypothetical. As a real-world analogue, recent work found that AI-extracted building footprints have entered OSM (Fila *et al*. 2025). While these are model-derived analytical outputs rather than generative images, their unmarked inclusion in open datasets similarly can increase coverage in data-sparse regions, they may also introduce errors – a model trained on these data might learn peculiar edge shapes or omit subtle features that human mappers would include, affecting tasks like urban change detection. Likewise, in the realm of geospatial textual corpora, Wikipedia entries (often used to enrich gazetteer database) are now frequently augmented or written by language models (Brooks *et al.* 2024). When unscreened AI-generated content is absorbed into training dataset, models inadvertently learn from synthetic artifacts rather than authentic signals. Our discussion echoes a worrying trend noted by Thompson *et al*. (2024): over half of publicly available web content in the near future could be AI-generated, meaning that any web-scraped geospatial dataset might carry a significant fraction of synthetic information by default.

The unsupervised injection of synthetic data into public datasets poses several ethical and practical problems. First, it undermines the validity of models trained on these data – performance metrics may be inflated if models effectively ‘overfit’ to synthetic patterns that superficially resemble reality but lack true complexity. Second, it raises issues of consent and transparency: users of open data typically assume the content is human-generated or at least quality-controlled. If AI-generated satellite images or location labels are mixed in without annotation, users cannot give informed consent to the risks, nor can they correct errors that stem from synthetic origins. Third, there is a compounding distortion effect. AI outputs tend to reinforce the patterns of their training data; if they dominate a dataset, they can drown out authentic minority signals. For example, if a GeoAI system generates many ‘average’ looking urban street images and those get added to a global dataset, unique local architectures or rare land uses might be systematically underrepresented. Over time, models trained on such a corpus could become blind to real but infrequent place-based features – effectively a form of data censorship caused by algorithmic over-representation of the norm.

Beyond these immediate risks, the broader ethical implications concern how the use of synthetic geospatial data transforms the authenticity of geographic knowledge and, by extension, its moral accountability. When AI-generated representations circulate through public datasets, the line between observation and simulation becomes increasingly blurred. What begins as a technical efficiency problem thus evolves into an ethical issue (Brey 2020): a question of how AI systems reshape what counts as ‘truthful’ or ‘trustworthy’ geographic information. The recursive reuse of synthetic imagery can detach GeoAI from empirical grounding, producing data that describe not the world as it is, but the world as previous models imagined it. This raises a responsibility that is both scientific and moral – to safeguard the integrity of geographic evidence in an era of automated data generation.

Place and digital placelessness

The recursive reliance on synthetic geospatial data also leads to a progressive abstraction of place. As the textures and specificities of local environments are blurred across generations, what is diminished is not merely visual detail but the contextual and affective particularity that gives place its meaning. 'Place' has long been understood not as a neutral spatial container but as a lived, relational, and emotional construct – rooted in experience, memory, and cultural distinctiveness (Relph 1976, Tuan 1977). The fading of such place-specific attributes in synthetic data generation therefore represents more than a technical limitation: it points to the ethical boundaries of applying AI to geographic knowledge (Floridi 2011).

When models recursively trained on synthetic data cease to recognize the specificities of place, they tend to produce a homogenized geography – one that smooths out difference and substitutes statistical coherence for lived complexity. This tendency echoes long-standing concerns about abstraction and de-contextualization in quantitative or data-driven representations (Ash *et al*. 2018). In the context of GeoAI, however, the process takes on a distinctly digital form. We describe this phenomenon as digital placelessness – a condition in which GeoAI systems, learning from their own outputs, generate spatial representations that are visually coherent yet ontologically detached from the embodied realities of place.

Unlike the 'placelessness' that Relph (1976) associated with architectural standardization and global modernity, digital placelessness emerges from algorithmic recursion and data abstraction. It transforms place from a lived, situated experience into a reproducible data pattern. What fades is not only the visual uniqueness of place, but also its epistemic and affective dimensions – the grounding of geographic knowledge in human and environmental multiplicity. Recognizing digital placelessness invites reflection on the broader limits of AI's universalizing tendencies and reaffirms the need for place-sensitive, contextually grounded approaches in GeoAI research and practice (Crawford 2021).

Transparency and provenance-aware curation

The above points converge on a key requirement for a healthy GeoAI ecosystem: transparency in data and model provenance. One of the clearest ethical implications of our study is the need for provenance-aware data curation. If we are to mitigate spatial errors and distortions, we must be able to trace what data went into a model and under what conditions that data were produced. Provenance tracking becomes especially crucial when synthetic data are involved. Without transparency, a model failure might only be discovered after its application in the real world. This is unacceptable for high-stakes geospatial applications.

Model transparency entails that developers openly document not just model architectures and accuracy metrics, but also the composition of training datasets – including the presence and proportion of synthetic data. We echo recent calls in the AI policy community for 'nutrition labels' or datasheets that describe the lineage of training data (Holland *et al*. 2020, Gebru *et al*. 2021) – an idea that directly supports GeoAI as well. For geospatial datasets, this could mean logging whether an image came from a camera or a generative model, or whether a map annotation was human-made or AI-suggested. Provenance metadata should accompany dataset releases and model publications so that other researchers can assess distortions and collapse risks. For example, if a segmentation dataset is, say, 30% composed of GAN-generated street scenes, one might interpret a model's performance with more caution, especially on fine-grained or rare place-based features. Transparent provenance would allow the community to identify when a model's surprising behavior might stem from synthetic training inputs rather than solely from model architecture flaws.

Transparency is also the antidote to spatial errors or distortions creeping in unnoticed. Spatial data are inherently varied across different regions and cultures. If synthetic data generation is disproportionately reflecting certain contexts (e.g., predominantly North American urban scenes because the generative model was trained on such), a model could exhibit strong geographical bias or contextual skew towards those regions. A provenance-aware approach would flag such issues early: for instance, GIScientists could ensure that if synthetic images are added to a dataset, they are generated in a geographically diverse manner and labeled as such. Moreover, transparent accounting of data sources enables provenance-based rebalancing – one could downweight or remove model training samples that are traced to an over-represented synthetic source to improve spatial representation. Without this knowledge, we risk unknowingly embedding systematic distortions into GeoAI systems and perpetuating a false sense of place. As Jacobsen (2023) discusses, the use of synthetic data often comes with implicit normative claims that technology can transcend representational biases by sheer volume or variation, yet it may simply mask these distortions under a veneer of 'neutral' algorithmic output. Only by being transparent about data origins can we hold such claims to scrutiny and ensure that errors or distortions of place-based features are actively identified and addressed.

Lastly, provenance-aware data curation is a practice that might require new tools and community norms. We may need conventions for tagging data (for example, a simple flag in metadata for *'synthetic_generation = pix2pix_v1'* or *'source_model = StableDiffusion2'*). Community-driven platforms could incorporate checks that prompt contributors to declare AI assistance. At the institutional level, journals and conferences could encourage or even mandate that GeoAI papers disclose any synthetic data usage in training or evaluation. These steps would collectively foster an environment where transparency is standard. With provenance information at hand, one can curate datasets more intelligently – perhaps excluding synthetic data when evaluating 'real-world performance' or, conversely, specifically using synthetic data in controlled ways to probe model robustness. While a comprehensive governance framework is beyond the scope of this paper, our findings underscore the urgent need for such collective reflection. We hope this discussion helps catalyze the development of shared community norms – including clarity about who should take responsibility for establishing and maintaining them. Ultimately, our aim is to encourage an open, field-wide dialogue – through future short papers, special issues, and conference panels – about how the GeoAI community can collaboratively articulate, debate, and institutionalize the norms that will guide responsible synthetic data practices in the years ahead.

Implications for evaluation and experimental design

Beyond data practices, our work carries implications for how we evaluate GeoAI models and design experiments to test their limits. The recursive training experiment we conducted can be seen as a stress test for model robustness. It reveals failure modes that would be invisible under conventional one-generation evaluations. Typically, a GeoAI model is trained on a dataset and evaluated on a holdout set of real images; if it performs well, we declare success. However, this standard evaluation paradigm does not capture how the model might perform when its outputs become inputs to future models – a scenario increasingly plausible in the era of model-assisted data generation. Our findings suggest that GeoAI evaluation frameworks should incorporate multi-generation or feedback-loop testing when synthetic data are part of the pipeline.

One practical approach is to include a recursive evaluation protocol: for instance, after training a model on real data, use it to generate a synthetic dataset (as we did with pix2pix) and then train a second model on that synthetic data. The performance of the second-generation model on a real-data test set is highly informative – it acts as an indicator of how much fidelity and diversity the first model's outputs retained. In our case, we saw significant drops in performance, especially on rare features, by the second and third generations. If instead we had observed minimal drop, it would suggest that the model's synthetic outputs were nearly as informative as the originals. Thus, reporting performance across generations could become a new facet of GeoAI model evaluation. This kind of evaluation would directly measure resilience to model collapse and could be presented alongside traditional metrics. For example, a table in a model benchmark could include not only accuracy on real test data, but also accuracy after one synthetic retraining cycle (perhaps even a 'degradation rate' metric). We believe this would incentivize GIScientists to design models that minimize collapse effects – e.g., by preserving distributional breadth.

Our experimental design also highlights the need for GeoAI-specific evaluation frameworks that account for spatial heterogeneity and real-world use-cases. GeoAI models often serve in decision-support roles where consistent performance across space and time is crucial. An evaluation framework might thus include tests for spatial generalization (does the model collapse more on data from region X vs. Y when using synthetic augmentation?) and for temporal stability (if a model were iteratively retrained over time with new data, some of which might be AI-generated, does its performance on past scenarios remain stable?). Incorporating these considerations could lead to protocols for longitudinal GeoAI evaluation, where a model's life-cycle performance is simulated rather than just a one-off snapshot. This is analogous to how we might test an autonomous system over a long-duration mission rather than just in a single static trial.

More broadly, longitudinal evaluation offers a way to understand GeoAI as a temporal process rather than a static artifact. Observing models over time – across successive retraining cycles and evolving data environments – can reveal how algorithmic systems drift, stabilize, or degrade in response to changing inputs. This temporal perspective bridges the technical and human-geographical dimensions of GeoAI by linking performance monitoring to questions of stability, adaptation, and responsibility. In practice, longitudinal evaluation could take the form of periodic re-benchmarking, temporal provenance tracking, or model life-cycle audits that document how behaviors shift as synthetic data accumulates. Conceptually, it also invites reflection on AI's temporality – how the epistemic and ethical implications of GeoAI unfold over time, requiring not only technical vigilance but sustained interpretive engagement from the research community.

Finally, reflecting on our experiment's broader significance, we see it as a call to establish benchmarks for responsible GeoAI. By this we mean evaluation frameworks should not only measure accuracy but also track indicators of ethical risk – for instance, the disproportionate performance drop on rare place-based features can be treated as a fairness metric. If one model exhibits a 5% drop on rare place-based features after a synthetic cycle and another exhibits 50%, this is salient information about ethical robustness that should factor into model selection. Likewise, evaluation could incorporate a provenance analysis step: checking if any of the training or test data might have synthetic origins (intentionally or not), and analyzing how that influences results. As the field progresses, we may even envision pre-deployment evaluation checklists that include 'synthetic data exposure' tests much like software undergoes stress testing.

Concluding remarks

In conclusion, this study offers empirical evidence that GeoAI systems are indeed vulnerable to *'model collapse'* when trained on AI-generated data – a scenario that is increasingly plausible in real-world settings. Using a *pix2pix* framework on street-level imagery, we demonstrated how successive generations of models – each trained on the synthetic output of the previous one – suffer cumulative performance degradation. Importantly, this degradation was most pronounced in the rarer place-based features, those very details that often carry significant importance for public safety and algorithmic fairness. These findings affirm the theoretical predictions of model collapse (Shumailov *et al*. 2024) and translate them into the geospatial context: as synthetic data proliferate, models can lose their grasp on the long tail of the distribution, becoming less reliable and less equitable in their outputs.

The ethical stakes of this 'GeoAI collapse' are high. On one hand, our results sound a warning that the uncritical, context-blind use of synthetic data can undermine the fairness, accountability, and reliability of GeoAI systems. On the other hand, we do not advocate an outright ban on synthetic data – rather, we argue for a cautious and context-aware approach. Synthetic data remain a valuable tool for GeoAI: they can bolster training sets where real data are scarce and help models generalize to rare or hypothetical scenarios. However, this must be done with eyes open to the risks. We urge practitioners to use synthetic geospatial data as a supplement, not a replacement. In practical terms, this means carefully curating synthetic datasets (with human oversight), validating model performance on real-world benchmarks (especially for minority features), and avoiding deep recursive reliance on generations of AI-produced data.

Our work also highlights the need for new governance norms in the GeoAI community. As generative AI becomes intertwined with geospatial data pipelines, questions of data governance move to the forefront. We see a pressing need for standards that ensure transparency about data provenance – for example, community guidelines or even regulations that require labeling of AI-generated content in public geospatial databases. The GeoAI community might establish norms akin to a ‘provenance disclosure’ principle, where researchers and data providers openly state how a dataset was created and what portion of it (if any) is synthetic. Governance could also manifest in the form of industrial best practices: organizations deploying GeoAI models should institute internal review processes to detect model collapse signs (such as unusual drops in performance on updated data) and to decide when retraining should incorporate fresh real-world data to re-ground the model. There may even be a role for external audits or certification – analogous to quality seals – for datasets and models that have been vetted for synthetic content and collapse resilience. As indicated, ensuring the long-term health of GeoAI will likely require both bottom-up efforts (researchers adhering to higher standards of documentation and curation) and top-down measures (policy interventions for transparency and accountability in AI training data).

In wrapping up, we reemphasize the hopeful perspective: GeoAI’s future can be bright and innovative if we learn to navigate the pitfalls of synthetic data wisely, certainly alongside other technical, ethical, and societal challenges. The notion of ‘GeoAI collapse’ introduced here is not a fatalistic prediction, but a call to action. By recognizing the warning signs early – performance loss in rare place-based features, narrowing model perceptions, creeping errors, distortions or biases – the community can implement safeguards and best practices to prevent collapse. A resilient GeoAI ecosystem might involve policies like a ‘human data baseline’ (ensuring that each new model generation is grounded with a core of real data), routine audits for synthetic data influence, and open repositories of real geospatial data to counterbalance the synthetic influx. Ultimately, building such an ecosystem is about reinforcing the connection between our models and the geographic reality they aim to represent. If left unchecked, recursive reliance on synthetic data risks not only technical degradation but also the drift toward digital placelessness – a condition where geographic representations become visually coherent yet detached from the lived, contextual realities of place. An ethically grounded GeoAI will be one that remains faithful to the richness of the real world, values transparency and accountability in its processes, and is prepared to adapt as new challenges (and opportunities) arise. By steering GeoAI in this direction, we can harness the benefits of synthetic geospatial data and generative models while safeguarding against their perils, ensuring that GeoAI continues to serve society in a fair, reliable, and trustworthy manner.

Acknowledgements

We sincerely thank the editor-in-chief, the guest editors of this special issue, and the three anonymous reviewers for their time, careful consideration, and constructive feedback.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Bo Zhao is a Professor in the Department of Geography at the University of Washington, Seattle. His research lies at the intersection of geospatial technologies and the humanities, with recent work focusing on the social implications of GeoAI and other emerging geospatial technologies, as well as fake geographies. He contributed to the conceptualization, main writing, and revisions of the paper.

Yue Lin is an Assistant Professor in the Department of Geography & Geographic Information Science at the University of Illinois Urbana-Champaign. Her research interests include spatial data science, location privacy, and the societal implications of algorithmic systems. She contributed to the conceptualization and the design and execution of the experiments for this paper.

ORCID

Bo Zhao http://orcid.org/0000-0002-7834-6672
Yue Lin http://orcid.org/0000-0001-8568-7734

Data and codes availability statement

The data, codes, and instructions that support the findings of this study are available with the identifier at the link: https://doi.org/10.6084/m9.figshare.29386874.

References

Alibani, M., Acito, N., and Corsini, G., 2024. Multispectral satellite image generation using StyleGAN3. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17, 4379–4391.

Amiri, H., Yang, R., and Züfle, A., 2024. GeoLife+: large-scale simulated trajectory datasets calibrated to the GeoLife Dataset. *In*: *Proceedings of the 7th ACM SIGSPATIAL International Workshop on GeoSpatial Simulation*, Atlanta, GA, USA. ACM, 25–28.

Ash, J., Kitchin, R., and Leszczynski, A., 2018. Digital turn, digital geographies? *Progress in Human Geography*, 42 (1), 25–43.

Brey, P.A.E., 2020. Recent developments in Dutch and European philosophy and ethics of technology. *Techné: Research in Philosophy and Technology*, 24 (4), 36–42.

Burden, J., *et al.*, 2024. Legal aspects of access to human-generated data and other essential inputs for AI Training. *University of Cambridge Faculty of Law Research Paper*, 9 (35).

Brooks, C., Eggert, S., and Peskoff, D., 2024. The rise of AI-generated content in Wikipedia. *In*: *Proceedings of the first workshop on advancing natural language processing for Wikipedia*, Miami, FL, USA. Association for Computational Linguistics, 67–79.

Cordts, M., *et al.*, 2016. The cityscapes dataset for semantic urban scene understanding. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, Nevada, USA. IEEE, 3213–3223.

Crawford, K., 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press.

Fila, M., Štampach, R., and Herfort, B., 2025. AI-generated buildings in OpenStreetMap: frequency of use and differences from non-AI-generated buildings. *International Journal of Digital Earth*, 18 (1), 1–23.

Floridi, L., 2011. *The philosophy of information*. Oxford, UK: Oxford University Press.

Foody, G.M., 2002. Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 80 (1), 185–201.

Gibney, E., 2024. AI models fed AI-generated data quickly spew nonsense. *Nature*, 632 (8023), 18–19.

Gebru, T., *et al.*, 2021. Datasheets for datasets. *Communications of the ACM*, 64 (12), 86–92.

Goodchild, M.F., 1992. Geographical information science. *International Journal of Geographical Information Systems*, 6 (1), 31–45.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69 (4), 211–221.

Guo, H., *et al.*, 2021. Deep building footprint update network: a semi-supervised method for updating existing building footprint from bi-temporal remote sensing images. *Remote Sensing of Environment*, 264 (February), 112589.

Guo, Y., *et al.*, 2024. The curious decline of linguistic diversity: training language models on synthetic text. *Findings of the Association for Computational Linguistics: NAACL 2024—Findings*, Mexico City, Mexico. Association for Computational Linguistics, 3589–3604.

Hataya, R., Bao, H., and Arai, H., 2023. Will large-scale generative models corrupt future datasets? *In*: *Proceedings of the IEEE international conference on computer vision*, Paris, France. IEEE, 20555–20565.

Heris, M.P., *et al.*, 2020. A rasterized building footprint dataset for the United States. *Scientific Data*, 7 (1), 245.

Holland, S., *et al.*, 2020. The dataset nutrition label. *Data Protection and Privacy*, 12 (12), 1.

Hu, Y., *et al.*, 2022. Enriching the metadata of map images: a deep learning approach with GIS-based data augmentation. *International Journal of Geographical Information Science*, 36 (4), 799–821.

Isola, P., *et al.*, 2017. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1125–1134).

Ito, K., *et al.*, 2024. Translating street view imagery to correct perspectives to enhance bikeability and walkability studies. *International Journal of Geographical Information Science*, 38 (12), 2514–2544.

Jacobsen, B.N., 2023. Machine learning and the politics of synthetic data. *Big Data & Society*, 10 (1), 1–12.

Jain, J., *et al.*, 2023. Oneformer: One transformer to rule universal image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2989–2998).

Kendall, G., and Teixeira da Silva, J.A., 2024. Risks of abuse of large language models, like ChatGPT, in scientific publishing: Authorship, predatory publishing, and paper mills. *Learned Publishing*, 37 (1), 55–62.

Kim, E.J., and Bansal, P., 2023. A deep generative model for feasible and diverse population synthesis. *Transportation Research Part C: Emerging Technologies*, 148, 104053.

Knight, S., Bart, Y., and Yang, M., 2023. Generative AI and user-generated content: evidence from online reviews. *In: Northeastern U. D'Amore-McKim School of Business Research Paper* (No. 4621982).

Li, P., *et al.*, 2025. Making AI Less' thirsty. *Communications of the ACM*, 68 (7), 54–61.

Li, W., *et al.*, 2024. GeoAI reproducibility and replicability: a computational and spatial perspective. *Annals of the American Association of Geographers*, 114 (9), 2085–2103.

Li, Z., *et al.*, 2021. Synthetic map generation to provide unlimited training data for historical map text detection. *In:* Proceedings of the 4th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, GeoAI 2021, 17–26.

Li, Z., and Ning, H., 2023. Autonomous GIS: the next-generation AI-powered GIS. *International Journal of Digital Earth*, 16 (2), 4668–4686.

Li, Z., *et al.*, 2025. Machine-learning-enabled spatial pattern mining: evaluating the impact of imperfect inputs. *International Journal of Geographical Information Science*, 39 (10), 2362–2400.

Lin, Y., 2023. Geo-indistinguishable masking: enhancing privacy protection in spatial point mapping. *Cartography and Geographic Information Science*, 50 (6), 608–623.

Lin, Y., and Xiao, N., 2023. Generating small areal synthetic microdata from public aggregated data using an optimization method. *The Professional Geographer*, 75 (6), 905–915.

Lin, Y., *et al.*, 2023. Creating building-level, three-dimensional digital models of historic urban neighborhoods from Sanborn Fire Insurance maps using machine learning. *PloS One*, 18 (6), e0286340.

Lin, Y., and Zhao, B., 2025. Posthuman cartography? Rethinking artificial intelligence, cartographic practices, and reflexivity. *Annals of the American Association of Geographers*, 115 (3), 499–512.

Manvi, R., *et al.*, 2024. Geollm: Extracting geospatial knowledge from large language models. *In: The twelfth international conference on learning representations*, Vienna, Austria. OpenReview.net, 1–17.

Marasinghe, R., *et al.*, 2024. Towards responsible urban geospatial AI: Insights from the white and grey literatures. *Journal of Geovisualization and Spatial Analysis*, 8 (2), 24.

Marwala, T., Fournier-Tombs, E., and Stinckwich, S., 2023. *The use of synthetic data to train AI models: Opportunities and risks for sustainable development*. UNU Policy Brief, No. 1, 2023. Tokyo, Japan: United Nations University. https://unu.edu/sites/default/files/2023-09/UNU-Policy-Brief_1-2023_The-Use-of-Synthetic-Data-to-Train-AI-Models.pdf [Accessed 25 December 2025]

Neuhold, G., *et al.*, 2017. The mapillary vistas dataset for semantic understanding of street scenes. In: *Proceedings of the IEEE international conference on computer vision*, Venice, Italy. IEEE Computer Society, 4990–4999.

Nikolenko, S.I., 2021. Synthetic data for deep learning. In *Springer optimization and its applications* Vol. 174. Cham, Switzerland: Springer.

Oluoch, C., 2024. Crossing boundaries: The ethics of AI and geographic information technologies. *ISPRS International Journal of Geo-Information*, 13 (3), 87.

Papyshev, G., and Yarime, M., 2021. Exploring city digital twins as policy tools: a task-based approach to generating synthetic data on urban mobility. *Data & Policy*, 3, 1–18.

Peterson, A.J., 2025. AI and the problem of knowledge collapse. *AI & Society*, 40 (5), 3249–3269.

Prédhumeau, M., and Manley, E., 2023. A synthetic population for agent-based modelling in Canada. *Scientific Data*, 10 (1), 148.

Relph, E., 1976. *Place and Placelessness*. London, UK: Pion.

Romano, A., 2025. Synthetic geospatial data and fake geography: A case study on the implications of AI-derived data in a data-intensive society. *Digital Geography and Society*, 8, 100108.

Shumailov, I., *et al.*, 2024. AI models collapse when trained on recursively generated data. *Nature*, 631 (8022), 755–759.

Thompson, B., *et al.*, 2024. A shocking amount of the web is machine translated: insights from multi-way parallelism. *In: Findings of the Association for Computational Linguistics ACL 2024*, Bangkok, Thailand. Association for Computational Linguistics, 1763–1775.

Tuan, Y.-F., 1977. *Space and place: the perspective of experience*. Minneapolis, MN: University of Minnesota Press.

Walker, M., and Winders, J., 2025. GeoAI and Political Geography. *In*: *GeoAI and human geography: the dawn of a new spatial intelligence era*, 181–191. Cham: Springer Nature Switzerland.

Wei, Y., and Tyson, G., 2024. Understanding the Impact of AI-Generated Content on Social Media: The Pixiv Case. *In*: *Proceedings of the 32nd ACM International Conference on Multimedia*, Melbourne, VIC, Australia. Association for Computing Machinery (ACM), 6813–6822.

Wenger, E., 2024. AI produces gibberish when trained on too much AI-generated data. *Nature*, 631 (8022), 742–743.

Wu, C., *et al.*, 2018. Check-in behaviour and spatio-temporal vibrancy: an exploratory analysis in Shenzhen, China. *Cities*, 77, 104–116.

Wu, H., *et al.*, 2018. Generating realistic synthetic population datasets. *ACM Transactions on Knowledge Discovery from Data*, 12 (4), 1–22.

Xing, X., *et al.*, 2025. On the caveats of AI autophagy. *Nature Machine Intelligence*, 7 (2), 172–180.

Xu, H., *et al.*, 2024. Leveraging generative AI for urban digital twins: a scoping review on the autonomous generation of urban data, scenarios, designs, and 3D city models for smart city advancement. *Urban Informatics*, 3 (1), 29.

Yamaguchi, S., and Fukuda, T., 2023. On the limitation of diffusion models for synthesizing training datasets. *In: NeurIPS 2023 SyntheticData4ML Workshop*, New Orleans, Louisiana, USA. OpenReview.net, 1–8. http://arxiv.org/abs/2311.13090

Zhang, Y., *et al.*, 2024a. MapGPT: an autonomous framework for mapping by integrating large language model and cartographic tools. *Cartography and Geographic Information Science*, 51 (6), 717–743.

Zhang, Y., *et al.*, 2024b. GeoGPT: an assistant for understanding and processing geospatial tasks. *International Journal of Applied Earth Observation and Geoinformation*, 131, 103976.

Zhao, B., 2022. Humanistic GIS: toward a research agenda. *Annals of the American Association of Geographers*, 112 (6), 1576–1592.

Zhu, Y., *et al.*, 2023. SynMob: creating high-fidelity synthetic GPS trajectory dataset for urban mobility analysis. *Advances in Neural Information Processing Systems*, 36, 22961–22977.