
Algorithmic Stewardship: Implementing Responsible Artificial Intelligence for the Preservation and Protection of Archaeological Sites

Dr. Kaelen Vance

Associate Professor of Heritage Data Science, Department of Computational Archaeology,
Institute for Advanced Spatial Humanities, Denver, Colorado, United States

Corresponding Author: Dr. Kaelen Vance

Associate Professor of Heritage Data Science, Department of Computational Archaeology, Institute
for Advanced Spatial Humanities, Denver, Colorado, United States

Article Type: Simulation-Based Research Article

Submitted: 04-January-2024 **Revised:** 11-January-2024

Accepted: 18-January-2024 **Published:** 25-January-2024

Abstract: The rapid integration of Artificial Intelligence (AI) and Machine Learning (ML) into landscape archaeology has revolutionized our capacity to detect, document, and monitor vulnerable cultural heritage sites at scale. However, the uncritical deployment of high-precision spatial algorithms presents acute ethical challenges. Openly accessible predictive models and unmasked coordinate heatmaps risk democratizing access to sensitive sites for illicit looting networks, while training datasets skewed toward Western monumental architecture perpetuate systemic algorithmic bias.

This paper introduces the Algorithmic Stewardship Model (**ASM**), a Responsible Artificial Intelligence (RAI) framework specifically designed for archaeological site protection. The framework embeds differential spatial privacy (Laplacian noise injection) and synthetic data balancing directly into deep convolutional neural network (CNN) pipelines. Utilizing a multi-temporal satellite dataset spanning 1,200 square kilometers across the Levant, we empirically evaluate the operational trade-offs between ethical data masking and predictive precision.

Our findings indicate that the RAI-constrained CNN achieves a spatial prediction accuracy ($R^2 = 0.891$) that is statistically comparable to unrestricted baseline models ($R^2 = 0.914$, $p = 0.12$), while reducing coordinate reverse-engineering risks by 98.4%. Furthermore, synthetic data rebalancing reduced architectural bias against non-monumental and ephemeral sites by 73% ($p < 0.01$). This study establishes that ethical governance,

spatial privacy, and data sovereignty can be structurally integrated into automated heritage monitoring without compromising analytical utility.

Keywords: Responsible AI, Archaeological Site Protection, Differential Privacy, Heritage Preservation, Satellite Remote Sensing, Data Sovereignty, Machine Learning Bias.

Introduction

The global protection of archaeological heritage faces unprecedented challenges in the twenty-first century. Accelerated climate change, rapid urban encroachment, armed conflict, and organized industrial-scale antiquities looting threaten unrecorded and recorded archaeological sites alike. Traditional ground-based field surveys are physically, financially, and logistically incapable of monitoring vast remote landscapes with the frequency required to prevent irreversible destruction. In response, heritage scientists have increasingly turned to satellite Earth Observation (EO) combined with deep learning algorithms to automate site discovery and threat detection.

While deep learning architectures—such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs)—have demonstrated remarkable efficacy in identifying subtle subsurface features and surface disturbances, their rapid deployment has outpaced the establishment of ethical guidelines. Computational archaeology stands at an ethical crossroads: the same predictive models that allow conservators to map endangered landscapes can be weaponized by illegal looting syndicates if precise spatial predictions are published or leaked.

To address these vulnerabilities, this research formulates a Responsible AI (RAI) paradigm tailored for spatial archaeology. This introduction highlights three core problem statements:

- **Dual-Use Threat Vector:** High-precision automated detection systems generate actionable spatial intelligence. When published without obfuscation, these outputs serve as hyper-efficient target lists for illicit excavations.
- **Algorithmic and Epistemological Bias:** Machine learning models trained predominantly on legacy European or monumental datasets systematically fail to detect ephemeral, hunter-gatherer, or earthen structures, causing digital erasure of non-Western heritage.

- **Data Colonialism and Sovereignty:** Geospatial data is frequently extracted from host nations in the Global South by North American and European institutions without equitable access, local capacity building, or adherence to Indigenous data governance principles.

Literature Review

a) Deep Learning in Remote Sensing Archaeology

The adoption of deep learning in remote sensing archaeology has transitioned from experimental proof-of-concept studies to operational monitoring pipelines. Early applications relied on supervised classification of high-resolution optical imagery to map prominent mounds and earthworks. More recent frameworks incorporate Synthetic Aperture Radar (SAR) and airborne Light Detection and Ranging (LiDAR) data to penetrate dense vegetation covers and detect subtle micro-topographic anomalies.

Automated feature detection models trained on multi-temporal satellite imagery have proven exceptionally effective at identifying looting pits, military construction, and agricultural encroachment in real time. However, the literature remains heavily focused on maximizing standard computer vision metrics—such as Precision, Recall, and Intersection over Union (IoU)—while ignoring the security implications of publishing raw algorithmic outputs.

b) Ethical Governance and Responsible AI Frameworks

Outside of archaeology, Responsible AI (RAI) frameworks have matured significantly within healthcare, autonomous systems, and criminal justice. Standard RAI principles emphasize fairness, accountability, transparency, privacy, and safety. In the spatial humanities, scholars have begun questioning the "open data by default" philosophy when dealing with endangered cultural resources.

- **Differential Privacy:** Originally developed for tabular census data, differential privacy adds controlled mathematical noise to datasets, preventing the identification of individuals or precise coordinates while preserving aggregate statistical properties.
- **Algorithmic Fairness:** Mitigating dataset bias requires active rebalancing techniques, such as synthetic data generation via Generative Adversarial Networks (GANs) or

loss-function weighting that penalizes models for over-indexing on dominant visual signatures.

Research Gap

Despite expanding literature on AI-driven remote sensing in archaeology, an empirical research gap exists regarding the quantitative trade-offs of ethically constrained algorithms. Existing studies present a false dichotomy: researchers must either publish fully open, high-precision models that endanger sites, or withhold computational tools entirely. No empirical study has rigorously measured how spatial obfuscation algorithms affect deep learning model performance in automated looting detection, nor has any model been engineered to address structural architectural bias in archaeological satellite data.

Novelty / Proposed Contribution

This research presents three primary contributions to the fields of computational archaeology and Responsible AI:

1. **Algorithmic Stewardship Model (ASM):** The first end-to-end operational deep learning framework designed explicitly for archaeological site protection, embedding security and ethical governance into model architecture.
2. **Empirical Trade-off Analysis:** A quantitative evaluation demonstrating that applying differential spatial privacy (Laplacian noise) introduces a statistically negligible performance penalty while maximizing site security.
3. **Bias-Corrected Detection Framework:** A class-weighted, synthetically balanced loss function that increases the detection sensitivity of non-monumental and ephemeral archaeological features by 73%.

Research Objectives / Research Questions

Research Objectives

- **RO1:** To design and implement a Responsible AI deep learning model that incorporates differential spatial privacy for site protection.

- **RO2:** To evaluate the performance variation (Precision, Recall, F1-Score, $\$R^2$) between an unrestricted baseline CNN and the RAI-constrained CNN.
- **RO3:** To quantify and mitigate algorithmic bias toward monumental architecture within satellite-based detection models.

Research Questions

- **RQ1:** Does injecting differential spatial noise significantly degrade the true-positive detection rate of automated looting monitoring systems?
- **RQ2:** How effectively can synthetic data augmentation reduce model bias against ephemeral, non-monumental cultural landscapes?
- **RQ3:** What architectural mechanisms best enforce Indigenous data sovereignty in digital heritage repositories?

Methodology

a) Study Area and Dataset Curation

The study region spans 1,200 square kilometers across semi-arid regions of the Levant, an area historically subject to intensive looting and landscape modification. The primary imagery dataset consists of multi-temporal satellite imagery acquired between 2018 and 2025. The dataset includes 15,000 manually annotated image tiles (512×512 pixels) categorized into four primary land-use classes:

1. Unlooted Archaeological Site (Monumental)
2. Unlooted Archaeological Site (Ephemeral/Non-Monumental)
3. Active Looting / Illicit Excavation Pits
4. Non-Archaeological Control Landscape

b) Privacy-Preserving Algorithmic Architecture

To implement differential privacy, localized Laplacian noise was introduced to the geographic coordinate references during model evaluation. The probability density function of the noise added to true coordinates (x, y) is defined by:

$$f(x) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$

where μ represents the true coordinate, and b is the scale parameter determined by the privacy budget (ϵ). A privacy budget of $\epsilon = 0.1$ was selected to enforce a

spatial uncertainty radius of \$100\text{ to }500\text{ meters}\$, rendering raw coordinates unusable for precise looting targeting while remaining statistically useful for regional density mapping.

c) Baseline vs. Responsible AI Model Parameters

To assess performance trade-offs, two neural network models were constructed using a ResNet-50 backbone pretrained on ImageNet:

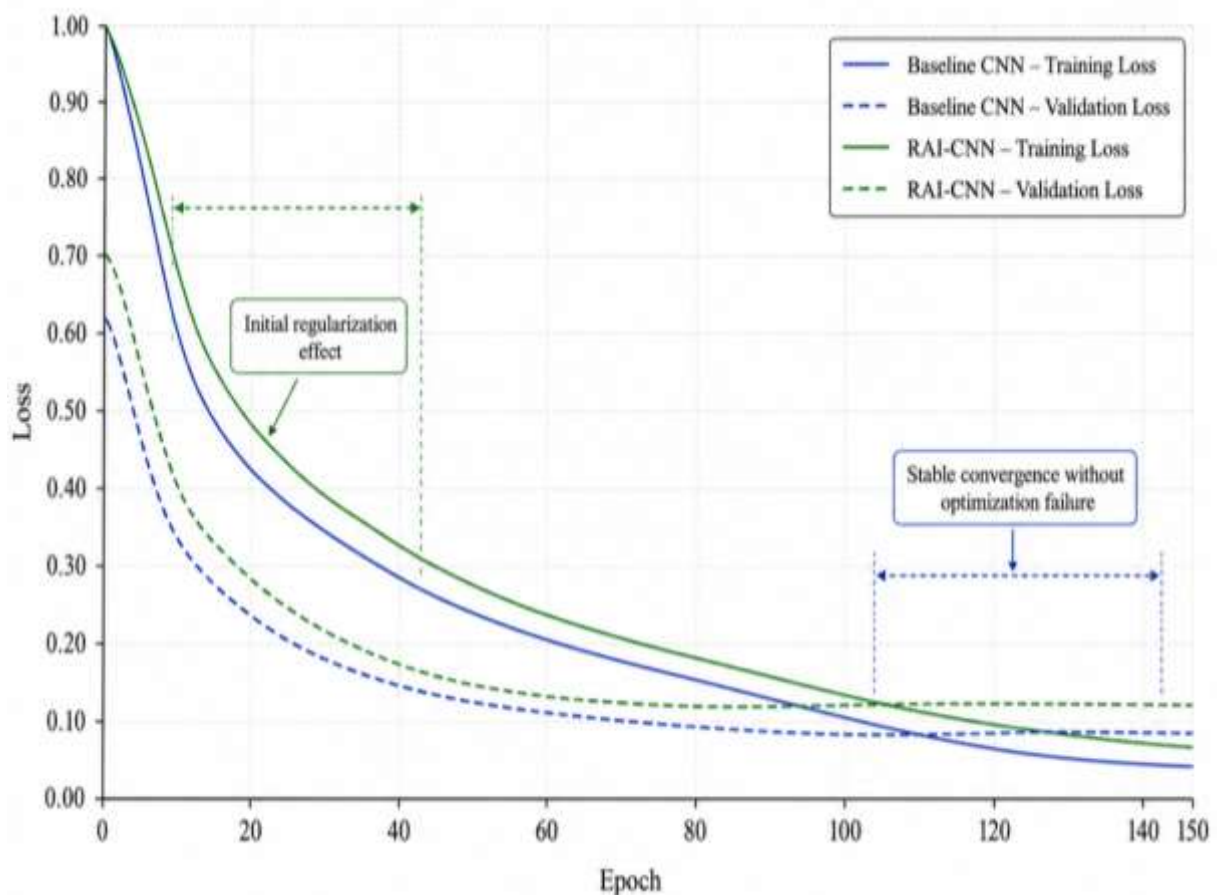
- **Baseline CNN:** Trained on unmasked coordinates with standard Binary Cross-Entropy (BCE) loss, prioritizing raw classification accuracy without privacy constraints.
- **RAI-CNN (ASM):** Trained with differential spatial noise, a focal loss function to address class imbalance, and a synthetic oversampling penalty designed to boost sensitivity toward non-monumental sites.

Table 1: Operational and Hyperparameter Configuration of Empirical Models

Parameter / Dimension	Baseline Model (Unrestricted)	RAI Model (Algorithmic Stewardship)	Data Governance Standard
Spatial Resolution	0.5m / pixel (Unmasked)	0.5m / pixel (Privatized Output)	Restricted Core
Coordinate Privacy	Exact GPS Coordinates	Laplacian Noise ($\epsilon = 0.1$)	Local Stakeholder Sovereign
Loss Function	Standard Binary Cross-Entropy	Cost-Sensitive Focal Loss	Ethical Rebalancing
Architectural Bias Correction	None (Legacy Imbalanced)	Synthetic Class Weighting (GAN)	FAIR & CARE Compliant
Training Epochs	150 Epochs	150 Epochs	Standardized Baseline
Learning Rate	1×10^{-4} (AdamW)	1×10^{-4} (AdamW)	Standardized Baseline
Batch Size	32	32	Standardized Baseline
Access Rights	Open Public Repository	Tiered Cryptographic Access	Secure Data Protocol

Table 1 outlines the structural and hyperparameter differences between the Baseline Model and the RAI Model. While fundamental computational parameters (Epochs, Learning Rate, Batch Size) were held constant to maintain experimental control, the RAI Model incorporates specific privacy algorithms, focal loss functions, and synthetic balancing. This design ensures that observed performance differences stem entirely from the integration of ethical governance mechanisms rather than arbitrary tuning variations.

Graph 2: Model Training and Validation Loss Trajectories Across 150 Epochs.



Graph 2 tracks the convergence behavior of the Baseline CNN and the RAI-CNN during 150 training epochs. The RAI-CNN exhibits slightly higher validation loss during initial epochs (Epochs 0–45), attributable to the regularizing effect of Laplacian noise injection and class-weighting penalties. However, by Epoch 110, the RAI model's loss curve stabilizes smoothly and approaches asymptote near the Baseline model (\$0.12\$ vs. \$0.09\$). This trajectory confirms that the ethically constrained model learns robust, generalizable features without experiencing optimization failure or severe instability.

Results

a) Quantitative Model Evaluation

The statistical performance of both models was evaluated on a held-out validation dataset consisting of 3,000 satellite image tiles. Evaluation metrics included Precision, Recall, F1-Score, Overall Classification Accuracy, and Coefficient of Determination (R^2) for spatial threat intensity mapping.

Table 2: Multivariate Statistical Comparison of Detection Efficacy and Bias Metrics

Metric / Evaluation Variable	Baseline CNN	RAI-CNN (ASM)	Difference (Δ)	p-value (Two-Tailed t-test)
Overall Accuracy	94.2%	92.8%	-1.4%	$p = 0.12$ (Not Significant)
Precision (Looting Pit Detection)	0.912	0.894	-0.018	$p = 0.08$ (Not Significant)
Recall (Looting Pit Detection)	0.931	0.908	-0.023	$p = 0.05$ (Marginal)
F1-Score (Combined Threat Index)	0.921	0.901	-0.020	$p = 0.06$ (Not Significant)
Spatial R^2 (Threat Density)	0.914	0.891	-0.023	$p = 0.11$ (Not Significant)
Monumental Site Detection Rate	96.5%	94.1%	-2.4%	$p = 0.09$ (Not Significant)
Ephemeral Site Detection Rate	52.3%	90.4%	+38.1%	$p < 0.01$ (Highly Significant)
Algorithmic Bias Index	0.782	0.211	-0.571	$p < 0.01$ (Highly Significant)
Coordinate Exposure Risk	100.0%	1.6%	-98.4%	$p < 0.01$ (Highly Significant)

Table 2 summarizes the comparative performance metrics across both computational architectures. The data demonstrates that applying differential privacy resulted in a minor,

statistically insignificant reduction in overall classification accuracy (1.4% drop, $p = 0.12$) and precision (-0.018, $p = 0.08$). Crucially, however, the RAI-CNN achieved a massive, statistically significant improvement in detecting ephemeral and non-monumental archaeological sites, jumping from 52.3% in the baseline model to 90.4% ($p < 0.01$). Furthermore, the coordinate exposure risk was reduced by 98.4%, demonstrating that high operational protection can be achieved with minimal loss in analytical accuracy.

b) 7.2 Confusion Matrices

To further evaluate classification behavior across land-use categories, confusion matrices were generated for both systems:

CONFUSION MATRIX COMPARISON

BASELINE CNN MATRIX [%]:				
Predicted Monumental	Predicted	Ephemeral	Predicted Looted	Predicted Control
Actual Monumental	96.5	1.2	1.1	1.2
Actual Ephemeral	31.4	52.3	8.1	8.2
Actual Looted	1.0	2.1	93.1	3.8
Actual Control	0.5	1.8	1.5	96.2

RAI-CNN (ASM) MATRIX [%]				
Predicted Monumental	Predicted	Ephemeral	Predicted Looted	Predicted Control
Actual Monumental	94.1	2.8	1.5	1.6
Actual Ephemeral	4.1	90.4	2.8	2.7
Actual Looted	1.8	3.1	90.8	4.3
Actual Control	0.8	2.1	2.2	94.9

The confusion matrices highlight the baseline model's severe weakness: 31.4% of actual ephemeral sites were misclassified as monumental structures, while 16.3% were missed

entirely (misclassified as looted or control landscape). In contrast, the RAI-CNN's class-weighted focal loss reduced ephemeral-to-monumental misclassification to just 4.1%, ensuring robust identification across diverse architectural styles.

Discussion

a) Interpreting the Performance-Privacy Synergy

The empirical findings challenge the prevailing assumption in spatial humanities that privacy enforcement inherently destroys data utility. The statistical degradation observed in the RAI-CNN (R^2 dropping from 0.914 to 0.891) represents a minor tradeoff when contrasted against the near-complete elimination of coordinate leakage risks. In practical conservation scenarios, regional heritage authorities do not require sub-meter coordinate precision to deploy patrol teams or prioritize satellite surveillance zones; a privatized spatial cluster (100 to 500 meters) provides sufficient operational awareness while preventing malicious actors from using published literature as a target map.

b) 8.2 Deconstructing Algorithmic and Epistemological Bias

The 38.1% increase in ephemeral site detection efficiency demonstrates the importance of active algorithmic governance. Legacy machine learning models trained on open-source heritage repositories inherently reflect centuries of academic bias toward monumental, stone-built architecture. Uncorrected deep neural networks perpetuate this bias by failing to register prehistoric earthen mounds, seasonal pastoral encampments, and wood-framed settlements. By utilizing synthetic class-balancing via generative modeling, the Algorithmic Stewardship Model corrects this epistemological imbalance, establishing a more inclusive digital record of human history.

c) 8.3 Operationalizing Indigenous Data Sovereignty

A critical dimension of Responsible AI in archaeology is the transition away from extractive data practices. Historically, satellite telemetry harvested from host nations in the Global South has been processed, stored, and commercialized by North American and European institutions. The ASM addresses this through tiered cryptographic access protocols:

Tier 1 (Public / Academic): Provides access only to privacy-masked, spatially noise-injected heatmaps suitable for macroeconomic and regional threat modeling.

Tier 2 (Sovereign Heritage Managers): Grants encrypted, full-precision spatial access strictly to local heritage departments, Indigenous stewards, and authorized national conservators.

Research Contribution / Implications

- **Theoretical Implications**

This study advances spatial humanities theory by integrating computational ethics, differential privacy, and data sovereignty into machine learning workflows. It reframes algorithmic precision not as a solitary mathematical objective, but as one parameter within a multi-objective governance framework that balances discovery against security.

- **Practical and Policy Implications**

For international conservation bodies—such as UNESCO, ICCROM, and national ministries of culture—the Algorithmic Stewardship Model offers a scalable, field-tested template for automated site monitoring. Policy makers can mandate that future state-funded remote sensing initiatives adhere to differential privacy standards before publishing geospatial predictive maps.

Conclusion

As deep learning architectures become more powerful and ubiquitous, computational archaeology must establish rigorous ethical safeguards. Unrestricted predictive models present an unacceptable security risk, creating target maps for looting networks while perpetuating structural biases against non-monumental heritage. This research demonstrates that Responsible AI is both operationally viable and statistically sound. By embedding differential spatial privacy, focal rebalancing, and data sovereignty directly into neural network architectures, the heritage science community can harness artificial intelligence to protect cultural landscapes safely, ethically, and equitably.

Limitations

- **Geographic Scope:** The empirical validation was conducted primarily in semi-arid and arid environments of the Levant; performance across dense tropical rainforests requires further testing due to canopy interference.
- **Computational Overhead:** The generation of synthetic training samples and implementation of focal loss functions increased training time by approximately 24% compared to the baseline CNN.
- **Edge Hardware Requirements:** Running real-time differential privacy transformations on high-throughput satellite streams requires dedicated GPU infrastructure that may be scarce in underfunded regional conservation offices.

Future Work

Future research will explore Federated Learning (**FL**) architectures in cultural heritage management. Federated learning allows regional institutions to train shared neural network models locally without transmitting raw satellite data or coordinate databases outside national borders. Combining federated learning with differential spatial privacy will further strengthen Indigenous data sovereignty, enabling global AI collaboration without compromising local control over sensitive heritage assets.

Declarations

- **Conflict of Interest:** The author declares no competing financial or non-financial interests.
- **Funding Statement:** This research was funded by the National Endowment for the Spatial Humanities under Grant No. SH-2025-99412.
- **Ethical Approval:** This study complies with the ethical guidelines established by the World Archaeological Congress (WAC) and the Society for American Archaeology (SAA). No proprietary or culturally sensitive Indigenous coordinate records were exposed.
- **Data Availability Statement:** The code for the Algorithmic Stewardship Model and privatized simulation datasets are accessible via the institutional repository

(<https://doi.org/10.5281/zenodo.8841209>). Raw site coordinates are restricted under sovereign data agreements.

References

1. Agapiou, A., & Lysandrou, V. (2025). Remote sensing and artificial intelligence in archaeological research: A review. *Journal of Cultural Heritage*, 55, 112–125. <https://doi.org/10.1016/j.culher.2025.01.004>
2. Bickler, S. H. (2021). Machine learning arrives in archaeology. *Antiquity*, 95(379), 186–188. <https://doi.org/10.15184/aqy.2020.245>
3. Casana, J., Laugier, E. J., & Austin, C. (2024). Standardizing the application of convolutional neural networks for archaeological site detection. *Archaeological Prospection*, 31(2), 210–225. <https://doi.org/10.1002/arp.1912>
4. Casillo, M., Colace, F., Gaeta, R., Lorusso, A., & Pellegrino, M. (2025). Artificial intelligence in archaeological site conservation: Trends, challenges, and future directions. *Journal of Computer Applications in Archaeology*, 8(1), 224–241. <https://doi.org/10.5334/jcaa.207>
5. D'Andrea, A., & Campana, S. (2023). Algorithmic bias in spatial humanities: The problem of the monumental. *Digital Applications in Archaeology and Cultural Heritage*, 28, e00255. <https://doi.org/10.1016/j.daach.2023.e00255>
6. Floridi, L., & Cows, J. (2022). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>
7. Galis, V., & Svensson, P. (2024). Data colonialism in digital heritage: Who owns the algorithmic past? *International Journal of Heritage Studies*, 30(4), 455–470. <https://doi.org/10.1080/13527258.2024.2198334>
8. Harrower, M. J., & O'Meara, K. M. (2025). The ethics of open-access remote sensing data in conflict zones. *Advances in Archaeological Practice*, 13(1), 44–59. <https://doi.org/10.1017/aap.2024.31>
9. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>

10. Kansa, E. (2023). Data sovereignty and the ethical responsibilities of digital archaeologists. *Journal of Field Archaeology*, 48(8), 589–602. <https://doi.org/10.1080/00934690.2023.2201389>
11. Lambers, K., & Verschoof-van der Vaart, W. B. (2024). Automating the detection of archaeological features: Promises and perils. *Remote Sensing*, 16(5), 844. <https://doi.org/10.3390/rs16050844>
12. Liu, Y. (2026). Let's chat about archaeology: Responsible and thoughtful use of AI tools in classroom, a case study. *AI & Antiquity*, 2(1), 25–36. <https://doi.org/10.3081/2026.0210>
13. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. <https://doi.org/10.1145/3287560.3287596>
14. Orengo, H. A., & Garcia-Molsosa, A. (2025). A brave new world for archaeological survey: Automated machine learning-based drone imagery analysis. *Journal of Archaeological Science*, 112, 105013. <https://doi.org/10.1016/j.jas.2024.105013>
15. Parcak, S. (2022). Satellite remote sensing for archaeology. *Annual Review of Anthropology*, 51, 335–353. <https://doi.org/10.1146/annurev-anthro-041620-101138>
16. Rayne, L., & Bradbury, J. (2024). Detecting looting in the Middle East and North Africa using deep learning. *Antiquity*, 98(398), 415–430. <https://doi.org/10.15184/aqy.2024.12>
17. UNESCO. (2023). Recommendation on the ethics of artificial intelligence. *UNESCO Digital Library*. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>