Article

AI, Cultural Heritage and Bias: Some Key Queries that Arise from the Use of GenAI

Anna Foka [*] and Gabriele Griffin

Article

AI, Cultural Heritage and Bias: Some Key Queries that Arise from the Use of GenAI

Anna Foka * and Gabriele Griffin

* Correspondence: anna.foka@abm.uu.se

Abstract: Our article AI, cultural heritage and bias examines the challenges and potential solutions for using machine learning to interpret and classify human memory and cultural heritage artifacts. We argue that bias is inherent in cultural heritage collections (CHCs) and their digital versions, and that AI pipelines may amplify this bias. We hypothesise that effective AI methods require vast, well-annotated datasets with structured metadata, which CHCs often lack due to diverse digitisation practices and limited interconnectivity. The paper discusses the definition of bias in CHCs and other datasets, exploring how it stems from training data and insufficient humanities expertise in generative platforms. We conclude that scholarship, guidelines, and policies on AI and CHCs should address bias as both inherent and augmented by AI technologies. We recommend implementing bias mitigation techniques throughout the process, from collection to curation, to support meaningful curation, embrace diversity, and cater to future heritage audiences.

Keywords: cultural heritage; GenAI; artificial intelligence; human-in-the-loop; bias

Introduction

Digital technology has drastically changed the possibilities for the curation and display of cultural heritage collections (CHCs). The physical and conceptual boundaries of such collections continue to expand, creating new opportunities for audiences to access and to engage with artefacts and cultural heritage (Geismar 2018). During cultural heritage collection digitalization processes the nuances of past heritage contexts need to be considered to ensure that cultures and diverse social groups are presented in an inclusive manner (Risam 2018). Heritage institutions traditionally use methods such as cataloguing and labelling to describe artefacts and to communicate such histories and cultures to the public. The fact that many collections were established through 'finds', excavations, expeditions, bought or seized by colonizers means that narratives related to colonization and oppression are inevitably part of analogue cultural records even if they are not made explicit within them. Cultural heritage is increasingly negotiated as a past practice that is (re)constructed in the present (Smith 2006: 3; Emerick 2014; Harrison 2013: 32, 165; Silverman, Waterton, and Watson 2017: 4- 8). Different dimensions of CHCs such as acquisition histories, museum history, ownership, location, the items themselves and curatorial guidance are all intertwined in creating an interactive system between people and information (Macdonald 2011). Critical heritage studies examines the nexus of people, heritage, and societal power in its challenge to conventional heritage discourses (Smith 2006: 281; Smith, Shackel, and Campbell 2012: 4). Heritage is thus a process 'understood as being produced through socio-political processes reflecting society's power structures' (Logan and Wijesuriya 2015: 569).

In digitizing CHCs, the question and answer protocol of new technologies such as ChatGPT or the production of synthetic images with Dall.e, MidJourney or Stable Diffusion immediately creates a situation in which human and machine exist in a cognitively productive relationship; the human describes, the machine renders. Generative AI, also known as GenAI or GAI, is an artificial intelligence technology that can generate text, images, or other data using generative models, often in response to prompts. It learns the patterns and structure of input training data to generate new

data with similar characteristics. GenAI and the synthetic data it produces has been examined from a number of perspectives. The aesthetics of AI and its impact on visual cultural practices have been extensively discussed (Manovich & Arielli 2022; Wasielewski 2023a). Understanding such computationally-aided creativity, there is a need for a deeper investigation of the socio-material complexity in implementing GenAI for cultural dissemination (Hayles 2017; Rettberg 2023).

The research question that underpins this paper is whether and how a machine can interpret and classify human memory and its artefacts in retrospect in an inclusive manner. Given the inevitable presence of bias in CHCs and in their digitized versions (Ciecko, 2020; Kizhner, 2021: 607-640; Foka et al., 2023: 815-825), this article aims to discuss the challenges that automation brings as well as provide solutions from beyond the cultural heritage sector. CHCs are normally quite diverse unless they are following some metadata standards as digitized historical collections are the result of legacy digitalisation. Further, there is a lack of interconnectivity/interoperability of digitised collections: not everything is online, or well annotated, or using the same softwares and that may be picked by a GenAI or an aggregator such as, for example, Google Arts and Culture.

Materials and Methods

In this article we draw on two kinds of source material to answer the research question: existing literature on bias mitigation in CHCs, and an experiment we conducted with image generation using a GenAI platform. These materials were analysed using semantic and visual culture analysis, with an emphasis on thematic interpretation. Due to the specificities of our materials and methods we do not separate results and discussion. Instead, we begin by discussing what bias is and how it is defined, by synthesizing scholarship on both CHCs and other datasets. We then discuss how bias relates to training data and lack of humanities expertise in contemporary generative platforms. We conclude that both scholarship as well as guidelines and policy on AI and CHCs should increasingly address bias as potentially augmented by AI technologies; measures should be taken from collection to data to curation to design AI and machine learning models to mitigate such bias to do justice to the inherent diversity and cultural complexity of collections.

Results and Discussion

What is bias and how does it leak into heritage datasets?

Bias as a concept is accompanied by ideas of prejudice, unfairness, distortion, and violation, inluding a systematic distortion of a statistical result due to a factor not allowed for in its derivation (Crawford, 2021). Bias may be as simple as an excluding description related to issues of race and ethnicity, age, gender, LGBTQIA+ communities, and ability. All CHCs involve selection, a form of bias in itself. Such selection is frequently accompanied by outdated descriptions of their artefacts that entail inclusion of some segments of society and exclusion of others, conforming to descriptions of a world very different from the contemporary. Dominant historical, national narratives and organisational legacies dictate what may be included and articulated in a collection (Lowenthal 2015; Smith 2006). While technology can, at least in theory, revolutionise how we understand the human contexts that CHCs carry and CHCs' 'democratisation' (Geismar 2018; Prescott and Hughes 2018), practice proves otherwise with the risk of carrying through biases of a not so distant past to the present and hence the future (Risam 2018; Thylstrup 2019; Wu 2020). As recently discussed in relation to newspaper archival collections, 'bias exists prior to any sampling…unbiased data—even as an idea—is essentially ahistorical data' (Beelen et al., 2023: 1-22).

Research into digital cultural data demonstrates how bias transitions from collections to datasets and then to platforms. Biases within museum collections can manifest in datasets, databases, and aggregators that increasingly employ AI technologies such as machine learning (Huster, 2013; Kizhner et al., 2021). Bias in CHCs entails issues of digital cultural colonialism and otherness, reflecting tensions between contrasting structures such as European/Western versus other, North versus South, and center versus periphery (Said, 1978; Sharp, 2002; Caton and Santos, 2008; Salazar, 2012; Risam, 2018). This also extends to gender. Kizhner et al. (2021) examine bias in the cultural heritage platform Google Arts and Culture, noting that the choices behind digitization, publication,

aggregation, and promotion often obscure institutional, social, and political circumscriptions. These perpetuate the status quo at scale (Kizhner et al., 2021). Kizhner et al. advocate making these epistemic choices transparent, documented, and interpretable. Davis et al. (2021) succinctly state that algorithms are animated by data, data come from people, people make up society, and society is unequal. Davis et al. (2021) discuss algorithmic reparation and intersectionality as frameworks to combat structural inequalities reflected and amplified by machine learning outcomes.

In computer vision, too, biases related to digital cultural colonialism and dominant epistemologies persist, leading to biased knowledge representations (Santos, 2018; Milan and Trere, 2019). To avoid merely replicating biases, AI technology must evolve to embrace complex, non-binary, and non-dominant interpretations. Critical perspectives from the humanities and social sciences play a vital role in highlighting these issues for more inclusive and equitable AI development practices. These perspectives emphasize the need for ethical AI development that addresses racial and gender discrimination, among other socio-ethical concerns.

Bias, especially racial and gender bias, extends across both technical and epistemological domains, with the gender binary serving as a deeply racialized tool of colonial control. The concept of auto-essentialization, recently introduced (Scheuerman, Pape, and Hanna 2021), describes how automated technologies reinforce identity distinctions rooted in colonial practices. The concept of auto-essentialization is explored through historical gender practices, particularly the establishment of the European gender binary via 19th and 20th-century disciplines such as sexology, physiognomy, and phrenology. These historical practices are viewed as predecessors to today's automated facial analysis technologies in computer vision. This connection underscores the necessity for a critical reassessment of AI/ML applications in image recognition, as they may represent modern iterations of longstanding technologically mediated ideologies (Scheuerman, Pape, and Hanna 2021; Rettberg, 2023: 118-127).

GenAI: An illustration of biased synthesis

Wasielewski (2023b: 71-82) examines the challenges faced by GenAI text-to-image generators like DALL·E and Stable Diffusion, focusing on their struggles with hand representation and object counting. While these tools have democratized AI-driven image creation, leading to a surge in creative outputs, they also exhibit significant limitations because they are mechanistic in their depiction of the objects, relying on pattern replication rather than contextual knowledge. This results in images that may appear superficially correct but lack nuanced understanding. The rise of generative AI models like ChatGPT and DALL·E has captured the public imagination; cultural and creative sectors increasingly turn to predictive models for analyzing and categorizing their materials (Berry & Dieter, 2015).

The opportunities GenAI affords are significantly structured by the CH sector that underlies them. As Griffin et al. (2023) have shown in the context of Sweden, a geographically large country with a small population (around 10.5 million) and a correspondingly small CH sector that is also quite fragmented, factors such as limited budgets, lack of AI expertise among CH staff, lack of professional mobility and of continuing professional training among CH staff, small collections, and no overarching national policy on the matter, can lead to scenarios where these factors are replayed in how AI is engaged with. This means that individual CHCs may acquire off-the-peg software solutions not trained on the data they are actually applied to, solutions that also lack interconnectivity and interoperability with softwares and systems in 'sister' CHCs, or may simply not (be able to) afford themselves of what AI and GenAI have to offer, thus isolating those CHCs both nationally and internationally.

While CHIs have traditionally been the domain of highly educated individuals, machines now play a significant role in evaluative tasks, with their effectiveness linked to data quality and categorization criteria. AI is in that sense able to reshape art and culture, blurring lines between authenticity and fabrication, especially in the era of advanced deepfakes. Machine learning, powered by extensive datasets which CHIs do not always have, enables the creation of synthetic images that possess a semblance of plausibility and authenticity, actively creating art and culture rather than merely documenting it. The application of deepfakes raises important ethical considerations,

particularly around trust and responsible use. It emphasizes the need for stakeholder engagement and participatory design approaches. AI-generated avatars offer new storytelling avenues for heritage enthusiasts and museum visitors, providing fresh perspectives on society, democracy, and humanity. The potential for misinformation in synthetic images is a growing concern. While generative models like DALL·E and Stable Diffusion can create images from text prompts, the interpretation and classification of these images often rely on algorithms trained on non-specialized datasets. The quality of these interpretations depends heavily on the data used and the collective human expertise in curating and preparing it.

AI may struggle to capture the nuanced characteristics of, say, Greek sculptures, such as their upright posture, detailed drapery, and iconic facial expressions. Achieving a satisfactory result often requires extensive human input, careful annotation of cultural heritage datasets, their curation, and fine-tuning, highlighting the ongoing need for human expertise in teaching AI tools high-level cultural competence. Take for example archaic kouroi, key to Greek art from 600-470 BCE, an idealized depiction of young men. These male figures exhibit a uniform appearance: nude, youthful, and muscular, especially in the chest and thighs. They stand upright with the left leg forward, arms at sides, and fists clenched. The face gazes straight ahead, featuring a rather formalistic enigmatic smile. Found across Greece as tomb markers or sanctuary dedications, kouroi show regional stylistic variations. They likely served as idealized representations of dedicants, the deceased, or even gods (Lorenz 2010: 133; for an example of a Kouros from Naxian Marble, ca. 590–580 BCE, see The Metropolitan Museum of Art collection, https://www.metmuseum.org/art/collection/search/253370).

We prompted a GenAI platform to create an archaic kouros and were faced with two completely different images: Figure 1 appears to be a female statue whereas Figure 2 is wearing some head gear that resembles a Corinthian helmet, a characteristic of classical warriors. The postures, formal features and even gender in Figure 1, are entirely off. None of these images correspond to the image or style of an authentic kouros. This indicates how tricky it (still) is to rely on GenAI to produce CH material without expert input.
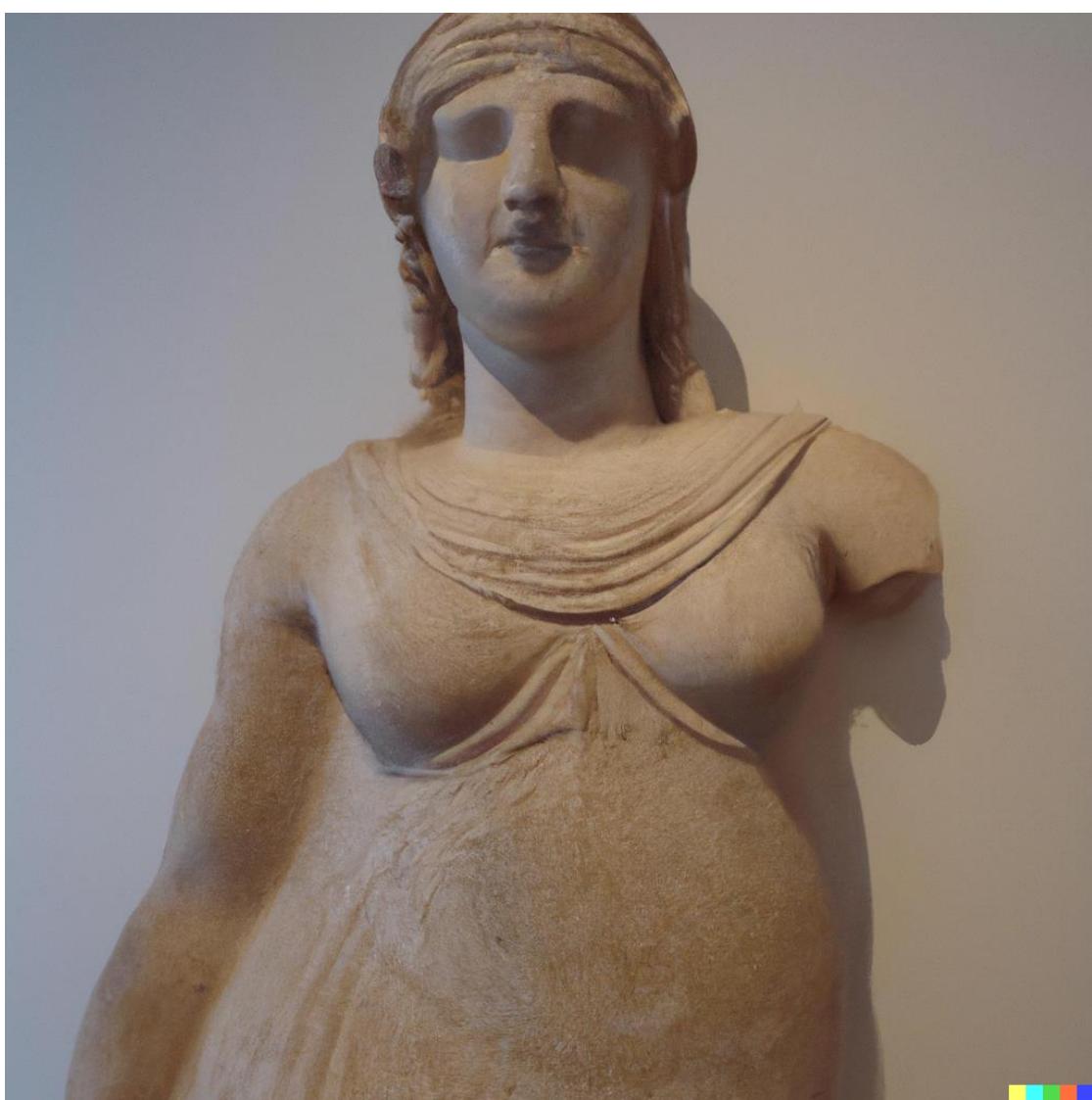
Figure 1. A GenAI kouros created with DALLE-2 using the prompt: a photorealistic photograph of an archaic kouros statue.

Figure 2. A GenAI kouros created with DALLE-2 using the prompt: 'a photorealistic photograph of an archaic kouros statue that resembles Apollo'.

At the time of writing this article there has been updates to DALL·E and DALL·E 2. The company states that they are no longer allowing new users for DALL·E 2. DALL·E 3 has higher quality images, improved prompt adherence, and they have started rolling out image editing - perhaps allowing for the possibility of customising images further.

When prompting an AI image generator such as DALL-E to create an image of an archaic kouros statue, the result may not fully capture the authentic form of the original sculpture. While it is possible to refine the output through iterative prompting and image variations, achieving a level of accuracy that would satisfy archaeological or classical art experts requires significant effort. More to the point, a novice user with little knowledge of what a kouros looks like, might create something completely inappropriate. Further, website users looking for such images would be misled regarding this kind of figure. The process of generating an 'authentic' representation hinges on significant expertise. For example, it would require training DALL-E 2 with expertly annotated archaeological datasets with a comprehensive understanding of the historical context, artistic techniques, and cultural significance. The ability to discern subtle details and stylistic nuances that define genuine Kouros sculptures is essential. Experts are therefore required to evaluate and select the most accurate AI-generated

images. In conclusion, while AI image generators can produce approximations of kouros statues, for example, achieving a level of accuracy that would satisfy scholarly standards remains heavily dependent on human expertise and intervention. The process of creating truly authentic representations requires a collaborative approach, combining the generative capabilities of AI with the specialized knowledge and discerning eye of human experts in the field of heritage.

Conclusion

Thoughtful application of AI in CHCs can provide crucial insights into heritage collections. To enhance their interpretive depth to a sophisticated level, we must develop AI systems capable of complex, nuanced analyses that avoid stereotypes. This evolution in image recognition technology is essential for unlocking the full potential of AI in understanding and in communicating CHC to the audiences of the future. Beyond this, we need both national policies and international agreements regarding interconnectivity and interoperability for CHIs and their collections, since the wherewithall to use AI and GenAI effectively in these institutions is not always readily available to individual institutions and their staff. At the same time, AI and GenAI are advancing rapidly and CHCs can find themselves left behind if they fail to engage with these new technologies.

References

1. Adams, M., Spence, J., Clark, S., Farr, J. R., Benford, S., & Tandavanitj, N. (2020). From sharing to gifting: A web app for deepening engagement. *Proceedings of EVA London 2020 30*, 48–49.
2. Anderson, S. (2020). Some provocations on the digital future of museums. In Winesmith, K. and Anderson, S. (eds), *The Digital Future of Museums: Conversations and Provocations*. Abingdon, New York: Routledge: 10–27.
3. Argyriou, L., Economou, D., & Bouki, V. (2020). Design methodology for 360 immersive video applications: the case study of a cultural heritage virtual tour. *Personal and Ubiquitous Computing*, 24(6): 843-859.
4. Bahrami, M., & Albadvi, A. (2023). Deep learning for identifying Iran's cultural heritage buildings in need of conservation using image classification and Grad-CAM. *arXiv preprint arXiv:2302.14354.*
5. Beelen, K., Lawrence, J., Wilson, D C S, & Beavan, D. (2023). Bias and representativeness in digitized newspaper collections: Introducing the environmental scan.   *Digital Scholarship in the Humanities*, 38 (1): 1–22.
6. Berg, H., Hall, S. M., Bhalgat, Y., Yang, W., Kirk, H. R., Shtedritski, A., & Bain, M. (2022). A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. *arXiv preprint arXiv:2203.11933.*
7. Berry, D. M., & Dieter, M. (2015). Thinking postdigital aesthetics: Art, computation and design. In *Postdigital aesthetics: Art, computation and design*. London: Palgrave Macmillan UK. pp. 1-11.
8. Caton K., & Santos C.A. (2008). Closing the hermeneutic circle photographic encounters with the other. *Annals of Tourism Research* 35: 7–26.
9. Davis, J. L., Williams, A., & Yang, M. W. (2021). Algorithmic reparation. *Big Data & Society*, 8(2), 20539517211044808: 1-12.
10. Emerick, K. (2014). *Conserving and managing ancient monuments: Heritage, democracy, and inclusion*. Woodbridge: Boydell & Brewer.
11. Griffin, G., Wennerström, E. and Foka A. (2023). AI and Swedish heritage institutions: Opportunities and challenges. *AI and Society*, DOI: 10.1007/s00146-023-01689-y.
12. Hayles, N.K. (2017). *Unthought: The Power of the Cognitive Nonconscious*. Chicago: University of Chicago Press.
13. Izsak, K., Terrier, A., Kreutzer, S., et al., (2022). Opportunities and challenges of artificial intelligence technologies for the cultural and creative sectors. Brussels: Publications Office of the European Union, European Commission, Directorate-General for Communications Networks, Content and Technology.
14. Fahse, T., Huber, V., & van Giffen, B. (2021). Managing bias in machine learning projects. In Ahlemann, F., Schütte, R. and Stiegllitz, S., *Innovation Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues*. London, Springer: 94-109.
15. Geismar, H. (2018). *Museum object lessons for the digital age*. London: UCL Press.

16. Hamilakis, Y. (2016). Decolonial archaeologies: from ethnoarchaeology to archaeological ethnography. *World Archaeology*, 48(5): 678-682.

17. Harrison, R. (2010). Introduction. In R. Harrison (ed.), *Understanding the politics of heritage*. Manchester: Manchester University Press: 5-42.

18. Harrison, R. (2013). *Heritage: Critical approaches*. New York: Routledge.

19. Huster, A. C. (2013). Assessing systematic bias in museum collections: A case study of spindle whorls. *Advances in Archaeological Practice*, 1(2): 77–90.

20. Kamiran, F., & Calders, T. (2010). Classification with no discrimination by preferential sampling. In *Informal Proceedings of the 19th Annual Machine Learning Conference of Belgium and The Netherlands,* Benelearn'10, Belgium, Leuven: 1-6.

21. Kizhner, I., Terras, M., Rumyantsev, M., Khokhlova, V., Demeshkova, E., Rudov, I., & Afanasieva, J., (2021). Digital cultural colonialism: measuring bias in aggregated digitized content held in Google Arts and Culture. *Digital Scholarship in the Humanities*, 36(3): 607–640.

22. Knell, S. (2010). Introduction. In Knell, S. J, Aronsson, P., & Amundsen, A. B. (eds), *National Museums: New Studies from Around the World*. London: Routledge: 3–29.

23. Logan, W., & Wijesuriya, G. (2015). The new heritage studies and education, training, and capacity-building. In W. Logan, M.N. Craith, U. Kocke (eds.), *A companion to heritage studies*. Malden, MA: John Wiley and Sons: 557-573.

24. Lorenz, K. (2010). "Dialectics at a standstill": archaic kouroi-cum-epigram as I-box. In M. Baumbach, A. Petrovic, & I. Petrovic (Eds.), Archaic and classical Greek epigram. Cambridge University Press: 131-148.

25. Macdonald, S. (Ed.). (2011). *A companion to museum studies*. Malden, MA: John Wiley & Sons.

26. Manovich, L. & Arielli, E. (2021). *Artificial Aesthetics, Manovich.net* (online)

27. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys* (CSUR), 54(6), 1-35.

28. Milan, S. & Trere, E. (2019). Big data from the south(s): beyond data universalism. *Television and New Media*, 20(4): 319–35.

29. Ooghe, B., Waasland, H. C., & Moreels, D. (2009). Analysing selection for digitisation. *D-Lib Magazine*, 15(9/10).

30. Pastaltzidis, I., Dimitriou, N., Quezada-Tavarez, K., Aidinlis, S., Marquenie, T., Gurzawska, A., & Tzovaras, D. (2022). Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*: 2302-2314.

31. Rettberg, J. W. (2023). *Machine Vision: How Algorithms are Changing the Way We See the World*. John Wiley & Sons.

32. Risam, R. (2018). Decolonizing the digital humanities in theory and practice. In Sayers, J. (ed.) *The Routledge companion to media studies and digital humanities*. London: Routledge: 78-86.

33. Salazar, N. B. (2012). Tourism imaginaries: a conceptual approach. *Annals of Tourism Research*, 39(2): 863–82.

34. Said, E. W. (1978). *Orientalism*. New York: Random House.

35. Sharp, J. P. (2002). Writing travel/travelling writing: Roland Barthes detours the Orient. *Environment and Planning D: Society and Space*, 20(2): 155–66.

36. Silverman, H., Waterton, E., & Watson S. (2017). An introduction to heritage in action. In H. Silverman, E. Waterton, & S. Watson (eds.), *Heritage in Action: Making the Past in the Present*, Cham: Springer: 3-18.

37. Smith, L. (2006). *Uses of heritage*. London: Routledge.

38. Smith, L., Shackel, P., & Campbell, G. (2011). Introduction: Class still matters. In L. Smith, P. Shackel, G. Campbell (eds.), *Heritage, Labour and the Working Classes*. London: Routledge: 1-16.

39. Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21),* ACM, New York, NY, USA: 1-9.

40. Tunbridge, J.E., & Ashworth, G.J. (1996). *Dissonant heritage: the management of the past as a resource in conflict*. Chichester: J. Wiley.

41. van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, 144: 93-106.

42. Wang, T., Zhao, J., Yatskar, M., Chang, K. W., & Ordonez, V. (2019). Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. *Proceedings of the IEEE/CVF International Conference on Computer Vision*: 5310-5319.

43. Wasielewski, A. (2023a). *Computational Formalism,* MIT Press.

44. Wasielewski, A. (2023b). 'Midjourney can't count': Questions of representation and meaning for text-to-image generators. *The Interdisciplinary Journal of Image Sciences*, 37(1), 71-82.

45. Wasielewski, A. (2023c). Authenticity and the Poor Image in the Age of Deep Learning. *photographies*, 16(2), 191-210.

46. Wells, J. C., & Lixinski, L. (2016). Heritage values and legal rules: Identification and treatment of the historic environment via an adaptive regulatory framework (part 1). *Journal of Cultural Heritage Management and Sustainable Development*, 6(3): 345-364.

47. Wells, J. C. (2015). In stakeholders we trust: Changing the ontological and epistemological orientation of built heritage assessment through participatory action research. In B. Szmygin (Ed.), *How to assess built heritage? Assumptions, methodologies, examples of heritage assessment systems*. Florence and Lublin: Romualdo Del Bianco Foundatione and Lublin University of Technology: 249-265.

48. Winter, T. (2013). Clarifying the critical in critical heritage studies. *International Journal of Heritage Studies*, 19(6): 532-545.

49. Witcomb, A., & Buckley K. (2013). Engaging with the future of 'critical heritage studies': looking back in order to look forward. *International Journal of Heritage Studies*, 19 (6): 562–578.

50. Zhang, Y., Zhang, Y., Halpern, B. M., Patel, T., & Scharenborg, O. (2022). Mitigating bias against non-native accents. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*: 3168-3172.

51. Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*: 335-340.