http://dx.doi.org/10.14236/ewic/EVA2024.4

Ethical AI and Museums: Challenges and new directions

Stefania Boiano
InvisibleStudio Ltd
London, UK
stefania.boiano@invisiblestudio.net

Ann Borda
The Alan Turing Institute and
University College London, UK
a.borda@ucl.ac.uk

Giuliano Gaia
InvisibleStudio Ltd
London, UK
giuliano.gaia@invisiblestudio.net

Guido Di Fraia
IULM University
Milan, Italy
guido.difraia@iulm.it

In this paper, we consider challenges and new directions in the use of Artificial intelligence in museums and particularly the need for supporting ethical frameworks. Such frameworks aid in the equitable and responsible adoption of technology and new forms of participation which can extend the role of museums as social good agents.

Artificial intelligence. Creative AI. Digital heritage. Ethics. Museum collections. Machine learning.

1. INTRODUCTION

Museums and other cultural heritage organisations have the potential to be relevant, socially-engaged and ethical spaces for intercultural dialogue. These organisations have traditionally been resilient places holding experiences accrued by human societies over time and across boundaries and worlds. The emergence of new technologies, namely Artificial intelligence (AI) and machine learning (ML) applications can assist in scaling the participation of communities and heritage stakeholders in supporting intercultural dialogue and inclusion.

Although there is no global definition of AI, computer scientist Alan Turing defined it as the science and engineering of making intelligent machines, especially intelligent computer programs (Turing 1950). Generally, the concept has evolved into a focus on simulating human intelligence by machines – e.g. an ability to build some kind of perception of knowledge that uses statistical methods to carry out tasks commonly associated with human intelligence and that assist or replace human decision-making in those tasks. Machine Learning (ML) is a subfield of AI which identifies patterns in data to include supporting classification, pattern recognition, prediction and the generation of text, sound and images (Leslie et al. 2022).

AI and ML are opening innovative ways to make museums more immersive and interactive. Museums and cultural organisations have been exploring the opportunities of AI, as well as the obstacles to their use and ethical implications (Villaespesa and Murphy 2021). According to a study of US and UK museums by Villaespesa and Murphy (2021), the application of ML to sort, analyse, and describe museum collections is a particular opportunity. These examples are reinforced in principle by regulatory framework advancements, e.g. Ethics Guidelines for Trustworthy AI (UNESCO 2021), and alongside developments in digital cultural heritage towards open knowledge systems and community participatory practices in narrative co-creation and decolonizing approaches.

Notwithstanding, there remain acknowledged gaps in our understanding of emerging and dynamic forms of knowledge production using AI, e.g. ChatGPT and *AI Time Machine*™, in which AI foundation models are often interculturally or semantically insensitive or misinformed to contexts of lived experience. Such challenges could be addressed in part through the insights gained in applying AI to cultural heritage knowledge systems in support of intercultural dialogue.

This endeavour simultaneously requires the fostering of new ethical frameworks which can more effectively account for the accumulation of such human experience, not least represented in knowledge co-production and in the decoding-encoding of intercultural agency in cultural heritage itself.

2. MUSEUMS AND AI

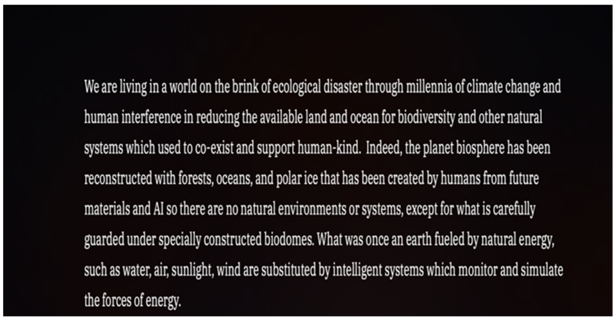
Museums have been piloting AI and natural language processing (NLP) enabled demonstrators for more than two decades (Abbatista et al. 2003; Boiano et al. 2003, Boiano et al. 2018, Borda and Bowen 2017). Chatbot applications, e.g. those made popular through Facebook Messenger, were quickly piloted by museums, most often as virtual guides (Gaia et al. 2019) when they appeared in 2015. For instance, the chatbot game developed by Invisible Studio for the *House Museums of Milan* project used Facebook Messenger to engage mainly younger visitors and teenagers in exploring four historic homes in Milan. In 2016, the Musee du quai Branly in Paris hosted *Berenson*, the robotic art critic, who interacted with visitors about their favourite and least favourite item in the collection, and through these interactions Berenson gradually built-up aesthetic preferences as it interacted with museum visitors (Styx 2023).

Among other potential applications are decoding unstructured knowledge embedded in cultural artefacts, context-based automated content creation and recommender systems, encoding context-specific cultural and personalised data. However, their actual implementation is currently limited due to the lack of resources and the inaccuracies created by algorithms. The role of crowdsourcing, and forms of online citizen science, have been filling this gap - that is the involvement of the general public in undertaking distributed tasks (such as tagging content, correcting text, etc.) using the Internet and various online computer-mediated communication platforms (Ceccaroni et al. 2023).

There is already an awareness of the potential of crowdsourcing in digital cultural heritage (Ridge et al. 2023) such as transcription efforts in the *Transcribe Bentham* project (Causer et al. 2018). The AI-enabled *MapReade*r application was developed by computational historians and curators, to help users to analyse large map collections of scanned and born-digital artefacts using deep learning and computer vision-based methods (Beelen et al. 2021). The industry involvement of Google DeepMind with classics researchers at Oxford University, the University of Venice, and the Athens University of Economics and Business supported the development of the AI application *Ithaca*, a deep neural network that can restore the missing text of damaged Greek text inscriptions, identify their original location, and help establish the date they were created (Assael et al. 2022). Harvard Art Museums (n.d.) is another example of a cultural institution piloting search and computer vision algorithms on its *AI Explorer* website. Users can choose an annotation search to find artworks which draws on tags, captions and object, as well as face and text recognition (Villaespesa and Murphy 2021).

An increase in creative AI tools available for museum practitioners will inevitably change the field in what is possible in content creation, curation and exhibition design. Text to image generators, such as DeepAI and DALL-E, are machine learning models that can create realistic and high-quality images from text. By inputting text descriptions of characters or environments, museum developers can rapidly generate visuals to incorporate in online or physical exhibitions or applications such as mobile games. Text to voice applications, such as *Murf.ai*, provide opportunities to create dialogue and lifelike voices for historical characters, for instance.

Generative AI systems can already create interactive narratives based on previously learned storylines and using text generation systems, such as the text-based fantasy simulation game *AI Dungeon* (n.d.). See Figure 1



We are living in a world on the brink of ecological disaster through millennia of climate change and human interference in reducing the available land and ocean for biodiversity and other natural systems which used to co-exist and support human-kind. Indeed, the planet biosphere has been reconstructed with forests, oceans, and polar ice that has been created by humans from future materials and AI so there are no natural environments or systems, except for what is carefully guarded under specially constructed biodomes. What was once an earth fueled by natural energy, such as water, air, sunlight, wind are substituted by intelligent systems which monitor and simulate the forces of energy.

Figure 1. *Author (AB) generated scene in AI Dungeon (https://aidungeon.com) using text prompts. Feb 2024.*

In a cultural heritage setting example, *Cultural Icons* is a web-based game experiment created by artist Gael Hugo, Artist in Residence at Google Arts & Culture Lab, that allows users to explore AI generated imagery, as well as engage in cultural conversations with Google's large language model PaLM2 (ai.google/discover/palm2/) and test their cultural knowledge through game play. Google Arts & Culture has concurrently established experimental sites developed by creative coders for piloting algorithms, such as X Degrees of Separation, which allows the user to select a random set of two cultural heritage images and discover their visual connection.

Among the challenges, however, concern the ownership of outputs of Generative AI which are built on others' source content. The latter has raised copyright issues and the difficulties in determining the ownership of the art and code generated by AI models (Murray 2023). In response, the content company Getty Images developed a proprietary AI tool in partnership with technology partner Nvidia to generate images from its immense database of digital media, thus avoiding copyright infringement (Villa 2023).

Similarly, there is the danger that AI will amplify already existing biases as it trains on existing data aggregated from vast and unvetted Internet and social media sources. Amongst the challenges for AI governance is that for algorithmic systems to process information about lived experience, these need to be synthesised into standardised data formats that can be interpreted and processed by machines. This process has given rise to ethical questions concerning the oversimplification of complex social phenomena leading to the erasures of minority identities, for example. AI practitioner and artist, Stephanie Dinkins, in residency at the Guggenheim Museum addresses marginalised groups "who are excessively affected by poor code design" (Segal 2023).

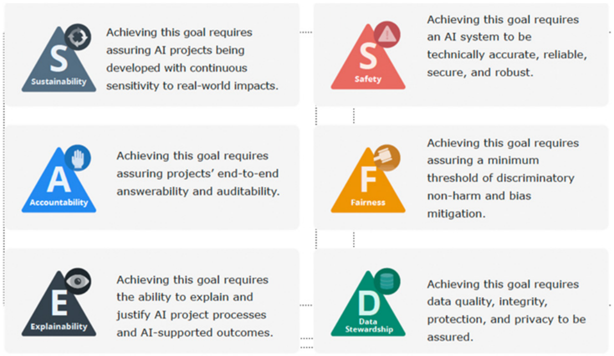
The limitations of Generative AI can be further understood in the example of *AI hallucination* when an AI model generates incorrect information but presents it as a fact. Curators at the Nasher Museum of Art at Duke University (Durham, North Carolina, USA) piloted the use of ChatGPT to curate an exhibition utilizing works of art from the Museum's collection and documented the results, including hallucinations (Merritt 2023). Amid a rapidly evolving landscape in which more AI-enabled systems are tracking with real-time geolocation data and using facial recognition, there is an equally urgent consideration in mitigating harmful risks through regulation and the impacts on vulnerable populations, such as children (Vosloo 2023). The implications for museums in this space is yet to be fully understood, but debates are ongoing as exemplified by the Misalignment Museum in San Francisco, USA (www.misalignmentmuseum.com) with its memorial exhibition to an imagined future in which AI has eradicated most of humanity.

3. ETHICAL FRAMEWORK FOR DIGITAL CULTURAL HERITAGE

To support an investigation of the multiple opportunities of AI use, we need to consider potential ethical issues of AI in museums by grounding perspectives in a theoretical framework.
As AI continues to evolve and transform society and the museum sector, it is imperative that ethical considerations are at the forefront of their construction. This is exemplified by the regulatory landscape and global directives in progressing appropriate ethical guidance such as the UNESCO *Recommendation on the Ethics of Artificial Intelligence* (2021), raising critical questions about the impact of this rapidly advancing technology on individuals and societies. The European Parliament has adopted resolutions addressing AI-related issues in education and culture sector in May 2021 devoting a separate section to the application of AI in the cultural heritage domain, including potential contribution to the preservation, restoration, on and management of the tangible and intangible cultural heritage (European Parliament 2021).

In 2019 the Museums + AI network engaged with 50 senior museum professionals, and leading academics across the UK and US to develop a toolkit (Murphy and Villaespesa 2020]. The network published a planning toolkit that allows museums to critically reflect on the capabilities and ethics of using AI within their collections. The toolkit is further outlined in the use case by IULM below. In brief, the toolkit helps a team to build an ethics workflow over the course of a project life cycle starting from the project goals to task-based phases, such as data input, data collection, data training, testing, application and evaluation. Individual museum organisations, for example, the Smithsonian Institution, have also worked towards establishing an AI values framework specific to their context (Dikow et al. 2023).

The *SAFE-D Principles* are another example of an ethical framework for evaluating more specifically the harms and benefits of data-driven technologies. Developed within the Public Policy Programme of the Alan Turing Institute (Leslie et al. 2022), the SAFE-D principles have been iteratively revised, tested, and validated with a wide-variety of stakeholders (Burr and Leslie 2022). See Figure 2.



Achieving this goal requires assuring AI projects being developed with continuous sensitivity to real-world impacts.

Achieving this goal requires an AI system to be technically accurate, reliable, secure, and robust.

Achieving this goal requires assuring projects' end-to-end answerability and auditability.

Achieving this goal requires assuring a minimum threshold of discriminatory non-harm and bias mitigation.

Achieving this goal requires the ability to explain and justify AI project processes and AI-supported outcomes.

Achieving this goal requires data quality, integrity, protection, and privacy to be assured.

Figure 2. *SAFE-D principles and icons (Leslie et al. 2023a).*

The acronym, 'SAFE-D' emphasizes 'safety' - an important component of trustworthy AI. The letters stand for the following five ethical principles:

- Sustainability
- Accountability
- Fairness
- Explainability
- Data stewardship

The SAFE-D principles provide high-level normative goals, - each of the SAFE-D principles has a subset of core attributes that help to specify and operationalise the principles throughout a project's lifecycle using a series of processes and activities. The core attributes serve as practical guardrails throughout a project's lifecycle.
text:

3.1 Sustainability

Sustainability requires the outputs of a project to be safe, secure, robust, and reliable. The sustainability of AI systems can rely on many factors, including the availability, relevance, and quality of data (Leslie et al. 2023b). For example, the technical sustainability of AI tools that are being deployed in museum contexts, require consideration in regard to how maintainable and reliable these tools can be in the long-term. Recent studies found that many AI systems are technically vulnerable (Zhang et al. 2024) and on-going maintenance efforts will be even more relevant as issues such as model degradation over time mean that AI-enhanced software require continuing efforts to remain accurate (Vela et al. 2022).

3.2 Accountability

Accountability requires transparency of processes and associated outcomes coupled with processes of clear communication that enable relevant stakeholders to understand how a project was conducted or why a specific decision was reached (Leslie 2019). This also concerns the question of how decisions are being made when it comes to the implementation of such tools. For example, images shared online have been appropriated for AI tools without explicit consent.

The Yahoo-Flickr Creative Commons 100 Million (YFCC100M) dataset contains 100 million media objects of photographs and video which carry a Creative Commons licence. The use of these images for AI has been questioned due to unanticipated re-use (Greshake Tzovaras and Ball 2019). Similarly, copyrighted artworks are being used for training AI for text-to-image applications without consent to create self-defence tools that modify the digital work to "poison" any tools being trained on them (Shan et al. 2023).

3.3 Fairness

Fairness determines whether the design, development, and deployment of data-driven technologies is fair which begins with recognising the full range of rights and interests likely to be affected by a particular system or practice, such as creating impermissible forms of discrimination (e.g. profiling of people based on protected characteristics or contributing to or exacerbating harmful stereotypes) (Leslie et al. 2023c). How fair an AI system depends heavily on the representativeness of the underlying data, as well as the quality of the collected data. In the absence of representative, high-quality data, AI are bound to reproduce and automate the biases in the underlying datasets used for training (EU 2022).

Connecting the issue of representation in museum projects, there is the question of whether projects that aim to deploy AI would benefit from data collection and whether the right kinds of data can be collected, as the use of data itself might be limited depending on how the project is rooted in dominant socio-technical contexts (Nafus 2023). The implementation of AI projects can further solidify participation biases, by privileging dominant forms of knowledge creation (Toupin 2024).

3.4 Explainability

Explainability refers to a property of an AI system to support or augment an individual's ability to explain the behaviour of the respective system (Burr and Leslie 2022). However, when the internal logic of a model is hidden or derives from empirical models that simply relate inputs and outputs, without explanations of the decision-making process or the internal functions, this is often termed as a 'black box' model (Hassija et al. 2024).

AI literacy is key to explainability so that individuals can become critical users of AI-enabled technologies. Broadly stated, AI literacy is the ability for individuals to understand, use, evaluate, and critically reflect on AI applications without the need to develop AI models themselves (Ng et al. 2021). For museums, the ability to locate, evaluate and use information (generated by humans and AI systems) critically and ethically is also essential for active participation and informed decision-making in an increasingly data-driven and algorithmic society (Kelley and Woodruff 2023).

3.5 Data stewardship

Data stewardship focuses on the data that undergirds AI and machine learning projects, including consideration of 'data quality' (e.g. whether the contents of a dataset are relevant to and representative of the domain and use context), 'data Integrity' (e.g. how a dataset evolves over the course of a project lifecycle) and legal obligations, including adherence to data privacy, protection and human rights compliance.

In considering data stewardship, museums have an opportunity to address the ethical and governance issues which this entails, not least due to its grounding in communities and data governance models (Micheli 2020). However, this critical stewardship role can be exacerbated by the evolving challenges of AI, such as image mislabelling by non-experts for training data resulting in them not being representative of source collections held by GLAM institutions (Dikow et al. 2023). Relatedly, there is a potential 'degrading' of the digital commons by generative AI models trained on publicly available data and public infrastructure but do not have mechanisms to reciprocate value captured to data producers or stewards (Huang and Siddarth 2023).

4. CASE STUDY: ITALIAN AI MUSEUM APPLICATIONS

In October 2023, InvisibleStudio, a cultural innovation studio based in London and Milan, organised a workshop on AI ethics in museums, in collaboration with Dr. Oonagh Murphy from Goldsmiths University and co-leader of the Museums+AI Network project. The event was held at the IULM AI Lab, a spinoff of IULM University in Milan, which focuses on AI in business and the humanities, founded and directed by Prof. Guido Di Fraia.

During the workshop, three AI case study applications in Italian museums were presented and examined from an ethical perspective. The three applications were selected as representative of three important directions in the development of AI in museums, specifically: artificial vision, virtual guides and generative AI. The three case studies are outlined below with analysis conducted using the *Museum AI Toolkit*, developed by the Museum+AI Network and freely distributed on https://themuseumsai.network/toolkit/. An Italian version of the toolkit is being developed by InvisibleStudio and IULM AI Lab in collaboration with Goldsmiths University and will be available on the Museum+AI Network website.

4.1 Artificial vision

The case study on artificial vision was presented by the Nemech/MICC research centre at the University of Florence. In this case, researchers developed the concept of an in-gallery game called *Strike-a-pose*. Strike-a-pose is a web application that analyses and evaluates human poses in comparison to those in famous paintings and statues (Donadio et al. 2022).

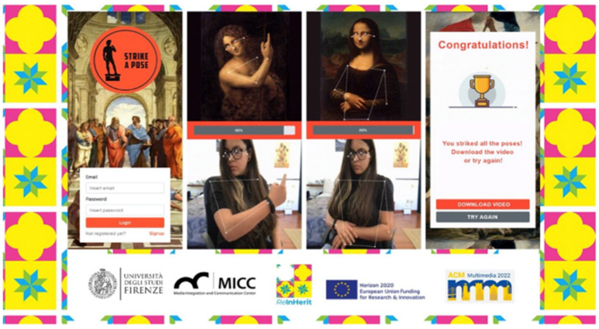

STRIKE A POSE
Email
Insert email
Password
Insert password
Login
Not registered yet?
Signup
Congratulations!
You striked all the poses! Download the video or try again!
DOWNLOAD VIDEO
TRY AGAIN
UNIVERSITÀ DEGLI STUDI FIRENZE
MICC
Media Integration and Communication Center
ReInHerit
Horizon 2020 European Union Funding for Research & Innovation
ACM Multimedia 2022

Figure 3. *Strike-a-Pose app screenshots 2022 (Courtesy of University of Florence)*

In Strike-a-pose, the user is challenged to reproduce in sequence the poses of some artworks from the museum's collections. Once all the poses have been matched, the application allows the user to generate a video that can be saved for any social sharing and provide information on the artworks. The video captures the user matching process and the overall interactive experience lived at the museum.

Artificial vision systems and expression recognition are focal points in the ethical debate on AI, reflected in their close monitoring under the European Union's proposed *AI Act*. It is thus crucial for the museum, both ethically and legally, to clarify the non-implementation of any automatic user identity recognition policies, nor the storage of biometric data. Given the use of externally developed recognition software, it is important that such software is free from pre-existing biases and capable of recognizing expressions across a diverse range of physical traits. Similarly, it is essential for apps like Strike-a-pose to function correctly with users who have various disabilities, such as the absence of certain limbs or use of a wheelchair.

As these types of apps use artwork images, it is vital for the museum to appropriately manage the usage and sharing rights of these images, whether owned by the museum or other institutions.

4.2 Virtual guides

The second case study focused on virtual guides, AI-powered characters capable of interacting with visitors and answering their questions. An example is *Nerobo*t, developed by Machineria and Ask Mona for the Rome Colosseo. Nerobot is a chatbot designed to provide practical information and can be utilised in two ways: either by clicking links or by posing open-ended questions, which are then analysed by the chatbot to match with pre-compiled answers provided by the museum staff. Nerobot is depicted in a comic style as Nero, the most famous of the Roman emperors, and introduces itself as such, albeit in an evidently ironic manner. Currently, it does not offer cultural content about the historical figure of Nero, except for a response about the Great Fire of Rome.

In museums, chatbots have been perceived as the institution's voice interacting with the public, necessitating accurate and reliable responses aligned with the museum's mission and values (Gaia et al. 2019). It is, therefore, essential to consider biases in the platform and biases represented in museum collections, in order to avoid a "double bias" effect (Murphy and Villaespesa 2020).

Regarding historical figures' portrayals, defining and understanding the limitations of a chatbot is key, especially when aiming to recreate the mindset of a character from another era. Informational chatbots are generally capable of answering basic queries, but ensuring privacy and security in conversations, particularly with third-party AI platforms, is essential.

The introduction of AI supported virtual guides can be seen as an evolution of widely-used audio guides and informational chatbots, however, replacing human guides in museums raises broader ethical concerns, including reduced human interaction during the museum visit.

4.3 Generative AI

The third case study highlighted a project that utilised ChatGPT at the National Cinema Museum in Turin, developed by Synesthesia. At the end of their visit, museum visitors were presented with the possibility to choose a few parameters to generate a script of a fictitious movie, such as genre, director style, time setting, and characters involved. Based on the user input, the system generated both a script and a movie poster, asking the user if they considered themselves as the "authors" of the script and movie, creating a connection with the screenwriters strike in Hollywood.

By putting generative AI systems in the hands of the public, the museum has a duty to prevent improper uses, and similarly to guarantee the privacy of the interactions and clarify any possible biases which might arise in the use of AI third party applications (in this case ChatGPT and Stable Diffusion). See Figure 4.

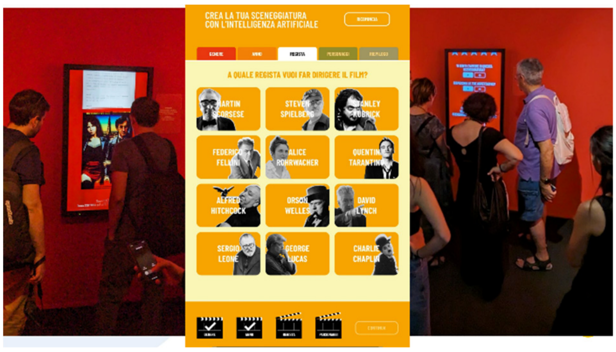


***Figure 4**. Images from the Cinema Museum installation 2023 (courtesy of Synesthesia)*

At the National Cinema Museum, a limited range of choices is offered to users to address the issue of AI-generated content potentially not aligning with the institution's mission and values. The challenge is that the vast creative potential of generative AI systems do not fully guarantee alignment. On the other hand, limiting options may reduce creative interactions and not address legitimate curiosities or interests of visitors. Balancing openness and systemic risk (e.g. privacy, security, bias) is a complex and evolving challenge in technology and knowledge production in museums.

5. CONCLUSION

Increasingly museums are identified as integral to a multi-stakeholder process, e.g. aiming at open data and participatory approaches to engagement (Godinho et al. 2019; Lesley 2019). The benefits of museum engagement can also be a key mechanism for knowledge co-creation, awareness raising, and behaviour change needed to operationalise ethical AI frameworks, including governance, safety and responsible use. To proactively address this agenda, museums can be engaged as a source of ethical AI literacy, contribute to trusted resources relating to responsible AI operationalisation, and co-lead in sociotechnical governance of responsible AI tools, among other opportunities.

Museums and cultural organizations are at a particularly significant juncture to re-imagine themselves filling in these ethical gaps as potential stewards of both the future and the past in which the digital citizen can participate in and equally contribute to closing those gaps arising in the evolving usages of AI in digital society.

6. CONTRIBUTIONS

AB and GG conceived the idea for the paper. AB, SB, GG drafted the paper. AB researched and wrote the introduction, museums and AI section and ethical framework section on SAFE-D principles. SB, GG, GDF conceived and wrote the Italian case study section.

7. ACKNOWLEDGEMENTS

AB acknowledges and thanks colleagues at the Alan Turing Institute for their expert contributions and support in reference to the SAFE-D principles, their development and application.

SB, GG, GDF thank IULM AI Lab for financing the Italian version of the Museums+AI Toolkit and Dr. Oonagh Murphy of Goldsmiths University for her precious contributions to the project.

8. REFERENCES

Abbattista, F., Bordoni, L. & Semeraro G. (2003) Artificial intelligence for cultural heritage and digital libraries, *Applied Artificial Intelligence*, 17:8-9, 681-686, DOI: 10.1080/713827258

Assael, Y., Sommerschield, T., Shillingford, B. *et al.* (2022). Restoring and attributing ancient texts using deep neural networks. *Nature* 603, 280–283. https://doi.org/10.1038/s41586-022-04448-z

Beelen, K., Hosseini, K., McDonough, K., Wilson, D. (2021). *MapReader: A Computer Vision Pipeline for the Semantic Exploration of Maps at Scale*. arXiv:2111.15592, *arXiv*, 30 Nov. 2021.

Boiano S., Caldarini M., Gaia G. (2003) Make Your Museum Talk: Natural Language Interfaces For Cultural Institutions, *Museums and the Web* Proceedings https://www.archimuse.com/mw2003/papers/gaia/gaia.html

Boiano S., Borda, A., Gaia, G. (2019) Participatory innovation and prototyping in the cultural sector: A case study. In: Weinel, J., Bowen, J.P., Diprose, G., Lambert, N. (eds) (2019) *EVA London 2019: Electronic Visualisation and the Arts*: pp. 18–26. DOI: 10.14236/ewic/EVA2019.3

Borda, A., Bowen, JP. (2020) Turing's Sunflowers: Public research and the role of museums. In: Weinel, J., Bowen, J.P., Diprose, G., Lambert, N. (eds) (2020) *EVA London 2020: Electronic Visualisation and the Arts*, pp. 32–39. DOI: 10.14236/ewic/EVA2020.5

Burr, C. and Leslie, D. (2022). Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. *AI and Ethics*, June 2022. doi:10.1007/s43681-022-00178-0.

Ceccaroni, L., et al. (2023). Advancing the productivity of science with citizen science and artificial intelligence. in *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research*, OECD Publishing, Paris, https://doi.org/10.1787/69563b12-en.

Cooper, C. B., Rasmussen, L.M. and Jones, E.D. (2022) *A Toolkit for Data Ethics in the Participatory Sciences*. Citizen Science Association. Available at: www.citizenscience.org/data-ethics

Dikow, R.B., DiPietro, C., Trizna, M.G., BredenbeckCorp, H., Bursell, M.G., Ekwealor, J.T.B., Hodel, R.G.J., Lopez, N., Mattingly, W.J.B., Munro, J., Naples, R.M., Oubre, C., Robarge, D., Snyder, S., Spillane, JL., Tomerlin, M.J., Villanueva, L.J., White, A.E. (2023) Developing responsible AI practices at the Smithsonian Institution. *Research Ideas and Outcomes* 9: e113334. https://doi.org/10.3897/rio.9.e113334

Donadio M., et al. (2022) Engaging Museum Visitors with Gamification of Body and Facial Expressions. In *MM '22: Proceedings of the 30th ACM International Conference on Multimedia* October 2022, 7000–7002. https://doi.org/10.1145/3503161.3547744

European Parliament. (2021). *Artificial intelligence in education, culture and the audiovisual sector*. Brussels. A9-0127/2021 https://www.europarl.europa.eu/doceo/document/TA-9-2021-0238_EN.html

Gaia, G., Boiano, S., Borda, A. (2019) Engaging Museum Visitors with AI: The Case of Chatbots. In: Giannini T., Bowen J. (eds) *Museums and Digital Culture*. Springer Series on Cultural Computing, pp 309-29. Springer, Cham. https://doi.org/10.1007/978-3-319-97457-6_15

Greshake Tzovaras, B., and Ball, M. P. (2019). Alternative personal data governance models. *MetaArxiv* https://doi.org/10.31222/osf.io/bthj7

Hassija, V., Chamola, V., Mahapatra, A. *et al.* (2024). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cogn Comput* 16, 45–74. https://doi.org/10.1007/s12559-023-10179-8

Huang, S., & Siddarth, D. (2023). Generative AI and the digital commons. *arXiv preprint arXiv:2303.11074*.

Kelley, P.G. and Woodruff, A. (2023). Advancing Explainability Through AI Literacy and Design Resources. *Interactions* 30, 5:, 34–38. https://doi.org/10.1145/3613249

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible* design and implementation of AI systems in the public sector. London: The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

Leslie, D., Rincón, C., Briggs, M., Perini, A., Jayadeva, S., Borda, A., Bennett, SJ. Burr, C., Aitken, M., Katell, M., Fischer, Wong, J., and Kherroubi Garcia, I. (2023a). *AI Ethics and Governance in Practice: An Introduction*. The Alan Turing Institute.

Leslie, D., Rincón, C., Briggs, M., Perini, A., Jayadeva, S., Borda, A., Bennett, SJ. Burr, C., Aitken, M., Katell, M., Fischer, Wong, J., and Kherroubi Garcia, I. (2023b). *AI Sustainability in Practice. Part One: Foundations for Sustainable AI Projects*. The Alan Turing Institute.

Leslie, D., Rincón, C., Briggs, M., Perini, A., Jayadeva, S., Borda, A., Bennett, SJ. Burr, C., Aitken, M., Katell, M., Fischer, C., Wong, J., and Kherroubi Garcia, I. (2023c). *AI Fairness in Practice*. The Alan Turing Institute.

Merrit, E. (2023). Curatorial Chatbot: An Experiment with AI at the Nasher Museum of Art. AAM Center for the Future Of Museums Blog Nov 28, 2023. https://www.aam-us.org/2023/11/28/curatorial-chatbot-an-experiment-with-ai-at-the-nasher-museum-of-art/

Micheli, M., Ponti, M., Craglia, M., & Berti Suman, A. (2020). Emerging models of data governance in the age of datafication. *Big Data & Society*, *7*(2). https://doi.org/10.1177/2053951720948087

Murphy O. and Villaespesa E. (2020). *AI: A Museum Planning Toolkit*. Goldsmiths University London. https://www.ukri.org/who-we-are/how-we-are-doing/research-outcomes-and-impact/ahrc/museums-artificial-intelligence-network/

Murray, M. D., Generative AI Art: Copyright Infringement and Fair Use (August 25, 2023). *SSRN*: https://ssrn.com/abstract=4483539 or http://dx.doi.org/10.2139/ssrn.4483539

Nafus, D. (2023). Unclearing the air: Data's unexpected limitations for environmental advocacy. *Social Studies of Science*, *0*(0). https://doi.org/10.1177/03063127231201169

Ridge, M., Ferriter, M., & Blickhan, S. (2023). *Recommendations, Challenges and Opportunities for the Future of Crowdsourcing in Cultural Heritage*: a White Paper. Digital Scholarship at the British Library. https://doi.org/10.21428/a5d7554f.2a84f94b

Segal N. (2023). *The Artist Preserving Histories with AI* | The Guggenheim Museums and Foundation. Available at: https://www.guggenheim.org/articles/checklist/the-artist-preserving-histories-with-ai

Shan, S., et al. (2023). Prompt-specific poisoning attacks on text-to-image generative models. *arXiv preprint arXiv:2310.13828*.

Styx L (2023) How are museums using artificial intelligence, and is AI the future of museums? *Museum Next* 18 June 2023. https://www.museumnext.com/article/artificial-intelligence-and-the-future-of-museums/ Accessed 14 December 2023.

Toupin, S. 2024. Shaping feminist artificial intelligence. *New Media & Society*, *26*(1), 580-595. https://doi.org/10.1177/14614448221150776

Turing, AM (1950). Computing machinery and intelligence. *Mind. New Series*, 59(236). (Oct. 1950), 433-460.

UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligenc*e. Office of the High Commissioner for Human Rights (OHCHR), United Nations, 23 November 2021. Available online: https://www.ohchr.org/sites/default/files/2022-03/UNESCO.pdf (Accessed on 12 April 2023).

UNESCO (2020). *UNESCO Report: Museums around the world in the face of COVID-19*. https://unesdoc.unesco.org/ark:/48223/pf0000373530

Vela, D., Sharp, A., Zhang, R. et al. (2022).Temporal quality degradation in AI models. *Sci Rep* 12, 11654. https://doi.org/10.1038/s41598-022-15245-z

Villa, A. (2023) Getty Releases AI-Image Maker Trained on Company's Data. *Art News*, 25 September 2023. Available at: https://www.artnews.com/art-news/news/getty-releases-ai-image-maker-trained-on-own-data-1234680408/

Villaespesa, E., Murphy, O. (2021). This is not an apple! Benefits and challenges of applying computer vision to museum collections. *Museum Management and Curatorship*, 36(4): 362-38 DOI: http://doi.org/10.1080/09647775.2021.1873827

Zhang, H., Ahmed, FA., Fatih, D., Kitessa, A., Alhanahnah M., Leitner, P., Ali-Eldin. A. (2022). Machine Learning Systems are Bloated and Vulnerable *arXiv preprint* arxiv:2212.09437