Original Research Article

Journal of Social Archaeology
2024, Vol. 0(0) 1–21

DOI: 10.1177/14696053241237949
journals.sagepub.com/home/jsa
Sage

(Re)framing built heritage through the machinic gaze

Vanicka Arora
University of Stirling, Stirling, United Kingdom of Great Britain and Northern Ireland

Liam Magee
Institute for Culture and Society, Western Sydney University, Penrith, NSW, Australia

Luke Munn
Research Fellow, Digital Cultures and Societies, University of Queensland, Saint Lucia, QLD, Australia

Abstract

Built heritage has been both subject and product of a gaze that has been sustained through moments of colonial fixation on ruins and monuments, technocratic examination and representation, and fetishisation by a global tourist industry. We argue that the recent proliferation of machine learning and vision technologies create new scopic regimes for heritage: storing and retrieving existing images from vast digital archives, and further imparting their own distortions upon this gaze. We introduce the term '*machinic gaze*' to conceptualise the reconfiguration of heritage representation via artificial intelligence (AI) models. To explore how this gaze reframes heritage, we deploy an image-text-image pipeline that reads, interprets, and resynthesizes images of several UNESCO World Heritage Sites. Employing two concepts from media studies—*heteroscopia* and *anamorphosis*—we describe the reoriented perspective that machine vision systems introduce. We propose that the machinic gaze highlights the artifice of the human gaze and its underlying assumptions and practices that combine to form established notions of heritage.

Corresponding author:
Vanicka Arora, University of Stirling, D24, Pathfoot Building, Stirling FK9 4LA, United Kingdom of Great Britain and Northern Ireland.
Email: vanicka.arora@stir.ac.uk

Keywords
Heritage photography, heritage gaze, machinic gaze, synthetic images, text-to-image models, generative AI

Introduction

Built heritage has a long, well-established relationship with visual representation, production, and consumption. Multiple scopic regimes have been in operation within the heritage industry and are continually evolving and diversifying, from careful artistic depictions of the romantic ruin, archaeological surveys, and cartographic representation through to photography, both as technical documentary evidence and as commercial tourist fantasy. Photography has long assisted in what Sterling (2019: 2) has termed the 'mythic representation of heritage as ideology', drawing attention to iconic or emblematic aspects of sites that reinforce narratives of power. The discussions around the 'gaze' in heritage have encompassed multiple ways of seeing. Chadha (2002: 380) suggests, for instance, with reference to the disciplinary project of archaeological photography in India, that multiple gazes are in operation simultaneously—the colonial, scientific, anthropological, and voyeuristic—while Wickstead (2009) considers the possibilities of moving beyond ideas of the male gaze and the Western gaze in archaeology, and instead approaching the gaze as diffused and ambiguous. For the purposes of our examination, however, we focus primarily on two established forms of viewing heritage—the tourist and expert gaze. Substantial work on heritage commodification and consumption builds on Urry's (1990) conceptualisation of the tourist gaze (see for instance Watson and Waterton, 2016; Waterton, 2009; Santos, 2016), while extending Foucauldian notions of the gaze for heritage are discussions around expert gaze (Bohrer, 2011; Moshenska, 2013; Smith, 2006; Winter, 2006).

Proliferating digital technologies and social media have intensified the heritage gaze and further complicated relationships between heritage and visual representation, especially in the context of photography. The introduction of machine learning technologies and generative artificial intelligence platforms that can now draw upon large archives of texts and images to resynthesise and produce 'photographs' in the absence of a 'real' object, temporality, or location are now positioned to substantially reconfigure these relationships.

We argue that the emergence of computer vision and, recently, of machine learning systems trained on image corpora reproduce both forms of the heritage gaze, alongside other styles and subjects, retaining as they do so existing social biases (Offert and Phan, 2022). However, this reproduction is not pure. In their reconstitution of synthetic photographs of heritage sites, image generating systems such as Midjourney adhere to conventions with palettes and perspectives, but also at times inject the uncanny differences of an alien observer or subject (Parisi, 2019). The differences between these machinic outputs and human expectation seem to belong to a novel, *sui generis* mode of visual perception and production, which we describe in this paper, following Denicolai (2021), as the 'machinic gaze'. 'Gaze' here serves a double purpose, referring to the

technical algorithms that make up computer vision and to the general 'way of seeing' that shares and yet is distinguished from human forms of apprehension. By directing computer vision algorithms to interpret and resynthesise a controlled archive of images, we offer a partial response to the question of what, in relation to heritage, of the human gaze, in its tourist and expert orientations, is *reproduced* by the machine, and what if anything is instead *introduced*? More generally: What does the machine see when it looks at heritage?

The expansive digitisation of vision has led to new possibilities in how machines consume and produce images (Azar et al., 2021). Social media image agglomerations have been systematised and organised into vast archives. With respect to these systems, two distinct kinds can be distinguished: *image-to-text* auto-captioning systems such as BLIP-2 (Li et al., 2023; Schuhmann et al., 2022; Zhang et al., 2023), and *text-to-image* generative systems such as Stable Diffusion, Midjourney, and DALL-E (Midjourney, 2022; Mostaque, 2022; Ramesh et al., 2022). We discuss the implications of both systems, though our focus is on the second, more novel system. With these generative AI models, the input of a text 'prompt', an instruction made up of typically English words that specify a subject, style, and format, generates synthetic images that, despite having no direct referent in their training sets or archives, can integrate parts of that prompt in often evocative and striking ways.

We focus on how this apprehension works to reproduce visual representations of heritage sites that have been subject to the explicit focus of both tourist and expert gaze. After a discussion of how to conceptualise the gaze, we describe experiments with machine-generated text and images, based on a small sample of images from UNESCO's World Heritage archive of sites. These experiments employ technical methods, using software libraries and machine learning systems, to read and decode these images into textual prompts, and then render those prompts as candidate reimaginings of the original images. We then comment on these machine-synthesised images and consider how these relate to both prompts and source images and conclude with implications of what the fast-moving field of machine learning might mean for the visual representation and production of heritage.

We undertook this exercise with three objectives. Our first and central objective is to consider the ways in which the image model captures, 'understands', and recreates the heritage site and the specific gaze directed towards these sites. The second is gaze-directed exploration of the politics of visual representation of global sites of heritage through the medium of the synthetically produced image. Properties of this synthesis, we argue, can condense and refract highly disparate human representations of heritage, marking out more clearly its own preoccupations and ideological attachments. The third objective is to begin to set out some of the parameters of the emergent relationships between heritage and synthetic photography. Using a 'textual' prompt to produce an image, we highlight the presence and endurance of the heritage gaze embedded in both text and image archives, mediated, and intensified through the machine. Our goal is not to assess the fidelity of auto-captioning or image generation systems or investigate these systems' capabilities to reproduce or extrapolate existing image archives. Rather, through our description of the machinic gaze, we hope to extend long-standing questions around the visual with respect to heritage—the heritage gaze, authenticity (or its absence), sense of place, and the

commodification of sites for tourist consumption in the context of emerging forms of generative AI.

Conceptualising the machinic gaze

The gaze, often with attached qualifiers ('male', 'colonial'), has an extensive history in heritage and adjacent fields of cultural studies (Wickstead, 2009). A common thread to distinct conceptualisations, from Mulvey's (2013) seminal essay on the male gaze to Urry and Larsen's (2011) discussion of the tourist gaze, is that *seeing* is never only a perceptual act, but is always informed by background assumptions, desires, prejudices, and power relations that inform interpretation of what is seen. Following work by media scholars (see Offert and Phan, 2022), we argue that despite the complexity of its datasets, training process, and software architecture, the machinic gaze as manifest in machine learning systems is similarly a social product. However, its relationship to diverse human gazes is not simply mimetic; rather it reproduces elements into representations that are often banal, and sometimes surreal and novel. While the trained machine has nothing to reference apart from its training set, at a certain scale and complexity mechanical *reproduction* can resemble an *introduction* of a novel palette, elements, and vision.

To conceptualise this process of transformation in the context of heritage, we begin with a discussion of two dominant modes of the heritage gaze: the tourist and the expert. Boundaries between the two are not always clear-cut, particularly now, as consumer devices and services make expert visions more accessible. Yet the dichotomy identifies imagistic qualities that help to account for certain aspects of the machine gaze, and to characterise what also distinguishes that gaze from dominant human vision paradigms.

In the context of archaeological monuments and sites, both tourist and expert ways of seeing have been further tied to forms of mechanical apprehension and capture since the inception of photography (Dicks, 2000; Sterling, 2016; Watson and Waterton, 2016). Shaped by a collectivised desire to witness scale and history, the polyvalent and complex tourist gaze (Urry and Larsen, 2011), alongside a supporting apparatus of travelogues, transport, and curation, has been stretched and magnified through the proliferation of social media platforms (Barauah, 2017; Oh, 2022). The expert gaze is similarly polyvalent, informed by disciplinary regimes ranging from archaeology and anthropology to architecture and conservation. The desire to document, authenticate, evaluate, and structure the object is central to this gaze, as is the construction of distance and objectivity (Beck and Sorensen, 2017; Bohrer, 2011; Wickstead, 2009). As other scholars have argued in relation to recent practice, this 'distance' is itself a multilayered phenomenon: one form of archaeological gaze reprises a positivist, scientific, and masculinist view of heritage observed, for instance, via top-down satellite imagery and GIS maps, while another—characterised as 'critical GIS' (Hacıgüzeller, 2012) or even 'gaze-critical' (Wickstead, 2009)—looks back reflexively on the techniques of archaeological production. In discussing the 'Europeanness' of heritage, Niklasson (2017) suggests a similar, more politically inflected distinction between past-preserving conservation and a present-oriented openness towards flexible interpretation. Across these distinctions, the expert orientation is still distinguished from that of the tourist by a precise and particularist

knowledge, which transfers to the preferred instruments, perspective, and types of attention directed toward heritage.

Similarly, the tourist gaze has been theorised as layered and multifocal. MacCannell (2001) draws upon the Lacanian conception of the gaze that stresses the effect of viewing upon the heritage spectator themselves, a move which recuperates the agency of the heritage observer. Viewing heritage does not simply involve a consuming tourist or calculative expert state but may effect a transformation of the spectator into a subject aware of their own historicity. The tourist experiences for example the strange sense of becoming an object for some other, future viewer or visitor—and as this object, also becomes a proper subject. Resisting efforts to subsume all touristic appreciation to that of cliche, Sterling (2016) has similarly argued that the seeing tourist is also an embodied figure, one who apprehends their own materiality in heritage encounters, and to varying degrees is also managed through deliberately arranged scaffolds and signs by heritage site managers. The body, in Sterling's account, in a certain sense anchors the otherwise cliched gaze within the singularity of the individual subject.

Both ways of seeing belong to a history of apprehension entwined with developments in optical technology (Kittler, 2010). Tourists, archaeologists, and other forms of expert viewing coordinate within networks of technical visuality: observing via a camera, decomposing an image, studying a map (Hacıgüzeller 2012; Sterling, 2019; Urry and Larsen, 2011), or constructing virtual and immersive environments (Champion, 2019; Forte, 2007). However, image-making AI systems do more than mediate, analyse, or mechanically reproduce (following Benjamin, 1986), and so seem to ask for a conceptual expansion of the machinic gaze. In generative systems, these patterns are mapped to words, so that when a prompt is submitted, the individual parts of the prompts serve as queries for finding these patterns; the patterns are then merged to produce a final image output. Data is however supplied as a closed set from which these patterns are learned. Unlike with photography or painting, there is no situated and embodied subject who encounters an object in what Crary (1990) terms the 'real' of human vision and perception. It is instead as though an image was produced by an artist forever trapped in a room, with only a captioned picture book for reference.

We propose that two specific operations of the machinic gaze can be identified through its reading and synthesising of images of built heritage. The first of these operations is *anamorphosis*. An old term of the pictorial arts describing the deformation of an object under different perspectives, anamorphosis was refreshed by Lacan (1998) to illustrate the distorting effects of unconscious desires on visual perception and cognition. In an analogous way, we describe the machine gaze as 'anamorphic' when it suggests unusual or bizarre affinities, provoked by what for a human viewer appear as accidental and unintentional, rather than essential properties of a source image or description. Extending this Lacanian connection, as MacCannell (2001) has earlier done in relation to the tourist gaze, anamorphosis also details the moment of reflexive human surprise at the realisation, in the face of machinic interpretation, of the contingency of their own ways of seeing. What appears first as technical error of translating prompt into image invites further questions as to how and why we perceive it *as* an error. In other words, this specific operation of the machinic gaze allows us to reflect upon our expectations of heritage

representation. Specifically, it allows us to interrogate our expectations of specific aesthetic values, forms, and style more closely in the outputs of generative AI. In the context of heritage, this offers up the possibility of querying specificities of the tourist and the expert gazes.

The second operation of the machinic gaze we attribute to the composite and synthetic character of generative AI systems like Midjourney and Stable Diffusion, which we describe as *heteroscopia,* a term coined by Jaireth (2000) in the context of Indian cinema. Jaireth gives heteroscopia two meanings: the first refers to a historical scopic regime or visual culture, while the second refers to how a given image may incorporate or reference other images, and so be more or less heteroscopic. Our own use adapts this second sense to the context of computer vision. In image generating systems, all outputs are essentially heteroscopic: they come from nowhere other than from an archive of existing images. The technical act of 'diffusion' in models like Midjourney and Stable Diffusion involves a twin process of adding noise to and subtracting it from an image corpus to learn to discriminate forms, styles, and colour compositions (Croitoru et al., 2023). These visual elements are related to captions in the corpus, and the training process produces in effect a network between visual elements and caption terms. Once the model is trained, prompts function as queries that in combination produce a synthetic image. This act of synthesis can sometimes reproduce a dominant gaze, and at others draw together disparate or incongruent elements into surrealistic montages or hybridised palimpsests. Heteroscopia here refers then to the extent and variation of gazes these systems render as outputs in response to prompts.

These two terms enable a move from the general technical operations of machines to a characterisation of the machinic gaze as applied to the heritage image—as something modelled on a codified and deracinated human vision that equally, as Parisi has noted (2019), apprehends its world through an uncanny and alien lens. The heteroscopic property allows for exploration and partial explanation of how the eventual image output appears as some composite of tourist, archaeological, and other forms of gazing—it describes the relation of the generated image to its inferred image sources. The anamorphic property captures instead the situation where the machine output traverses human conventions and expectations in the relation of image to text. When directed towards heritage, the machinic gaze reveals, at different moments, a dominant heritage scopic regime as well as moments of divergence that elicit opportunities for re-engagement.

Methods

In this section we explore this conceptual understanding of the machinic gaze through production of a small dataset of synthetically generated images. The dataset was developed to enable contrast across three dimensions: (1) publicly available image models; (2) visually distinctive heritage sites; and (3) expert-authored text and text derived from analysis of source photographs. Though specific to our study here, aspects of the approach outlined below suggest other uses in archaeological research, from auto-captioning to novel forms of image archive analysis. The machine, as we note, pays attention to what is

presented in images and texts differently, and while our purpose here is primarily to study that difference itself, we also acknowledge it can complement and correct the researcher gaze. To that end, we include in the Appendix links to code and datasets to allow replication and further exploration.

We produced the dataset in a sequence (Figure 1). First, we selected digital photographs from the UNESCO World Heritage Sites online archive. We then ran these images with three algorithmic interpretations (BLIP-2, Google Vision API, image EXIF metadata) to assemble a brief textual description for each image. These assembled descriptions were submitted in turn to three image generation systems in the form of textual 'prompts' (Stable Diffusion, Realistic Vision, and Midjourney) to produce a series of image samples—120 in total. Finally, we interpreted these images in terms of subject, composition, and deviations from the source images. We briefly discuss each of these steps below.

Image selection

We used the archive of photographs from the official website of UNESCO's World Heritage Sites as base images. Most of these photographs were taken by experts appointed directly by UNESCO's World Heritage Programme Office or by individual State Parties and are intended to simultaneously serve as official visual documentation of the site and ostensibly communicate a sense of its 'outstanding universal value' for a general audience In order to limit our search, we filtered images on two conditions: inclusion in the UNESCO World Heritage in Danger list and meeting criterion (iv), 'to be an outstanding example of a type of building, architectural or technological ensemble or landscape which illustrates (a) significant stage(s) in human history' (UNESCO, 2008). Of the 31 results returned, we then chose single Creative Commons-licensed photographs of five sites that contrasted with each other with respect to photo range and perspective, building typology, geographic region, historical style, and site description. This selection of sites was intended to highlight variations in the operationalisation of the heritage gaze and is not related to the sites' individual histories or World Heritage trajectories.

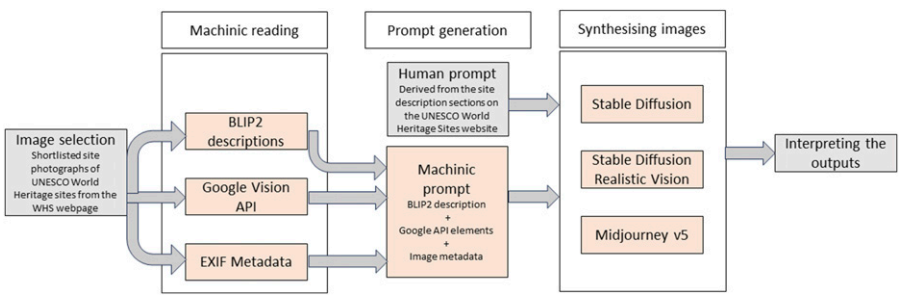

Figure 1. Machinic reading, prompt generation, and image synthesis pipeline.

Expert and machinic readings of heritage sites

We employed two techniques to produce prompts. The first takes the UNESCO-supplied description as an 'expert' view of the site. The second applies three computational techniques to extract information from the selected site photos, and combines these into a synthetic, automated prompt. Technique one uses both the visual and textual cues reflecting the expert gaze, reflected in UNESCO's WHS descriptions. The selected photographs are taken by UNESCO-appointed experts and a related group of expert group authors. Both privilege expert gaze of the sites themselves, attentive to what is most salient, distinguishing, or 'outstanding,' and what therefore must be compared against a repertoire of other built forms and sites.

In technique two, we first extract captions via BLIP-2, a vision-language based model which uses pre-trained image encoders and large language models to extract image captions (Li et al., 2023). The generated captions are short and generic, often five to 10 words. To enrich the prompt text further, we combined the caption with comma-delimited labels extracted from Google Vision's Application Programming Interface (API). These labels included computed image properties such as dominant colours, objects, locations, architectural features, geometric shapes, natural features, and colour schema, as well as individual objects. Each label includes a relevance probability, and we included the top 30 labels. In many cases, labels were redundant, misidentifying, or overly specific, and we pruned the list manually. Finally, we added metadata extracted from the digital image itself, including the type of camera, focal length, exposure time, and use of flash.

Machinic synthesis and its interpretation

The prompts or instructions produced through both methods were then submitted to three image generating systems. We selected Stable Diffusion and Midjourney, two systems widely discussed in 2023. Stable Diffusion is a text-to-image model that has been made open source by its developer, StabilityAI, and can be downloaded and operated on consumer devices. Midjourney is a service-based system that requires a subscription to operate, via the Discord social media platform. Both perform similar functions, converting a natural language prompt into one or more images that aim to 'represent' that prompt in a meaningful way. We used the latest versions of these two systems at time of writing: version XL in the case of Stable Diffusion and version 5.2 in the case of Midjourney. Stable Diffusion models can be adapted or 'fine-tuned' on much smaller data sets of images to produce styles or aesthetics. For further contrast we used an older version of Stable Diffusion (version 1.5) fine-tuned to generate photorealistic images, in a model named 'Realistic Vision'.

For each of the five sites, we applied the prompts generated through the approach described above to each of the three systems. We specified each system to generate four images for each prompt, producing a data set of 60 images (five sites x three systems x four images). For comparison, we also applied the UNESCO-supplied description for the selected site as a prompt to the same combination of site, system, and image variations, doubling the size of our data set to 120.[1]

Finally, we interpreted these sets of images in terms of their composition, form, subject selection, framing, colour palette, and aesthetic style. This interpretation, as we reflect upon in our findings and discussion, involves reflection upon the acts of seeing and reading of machine-generated images. It builds necessarily upon our own backgrounds in heritage and media studies, and consequently involves a specific form of what has been theorised as the expert gaze. Despite the limits of such interpretation, we look to avoid a specific judgement upon these machinic productions in terms of their approximation to some notion of 'ground truth' or as a quantitative exploration of bias within the underlying datasets of these systems (Salvaggio, 2022), instead focusing on unusual objects and style elements.

The machine imagines heritage

We discuss here three sets of images that contrast internally (across models and prompt) and externally (across sites). We use 'H-M' and 'M-M' to distinguish human-prompt-machine-generated from machine-prompt-machine-generated images.

Old towns of Djenné

Figure 2 shows a mid-distance elevational aspect of the mosque of Djenné, which is one the key structures identified in the description of the World Heritage Site. The adobe mosque appears in multiple photographs of the site, as one of the distinctive architectural landmarks within the urban ensemble of Djenné. The photograph frames the mosque tightly, editing out the immediate context of the marketplace or townscape that surrounds the mosque.

Figure 3 shows results of the 'H-M' process: four outputs (in rows) of three image models (in columns), in response to the prompt that was extracted from the description of Djenné on the UNESCO World Heritage Site website, which included phrases like 'typical African city', 'intensive and remarkable use of earth', 'mosque of great monumental and religious value' (UNESCO, 2023). In the case of Stable Diffusion (both versions), while the colour scheme of the UNESCO image is retained, no version of the mosque is produced in any of the images. SDXL (left column) shows, at different resolutions, a grid-like configuration of mud brick structures that approximates the sub-Saharan vernacular, but without the specificities of Djenné's architectural proportions or ornamentation. The Realistic Vision outputs (centre column) produce an approximation of a generic sub-Saharan settlement, small adobe buildings with thatched roofs—neither characteristic of the mosque nor of the general town. Midjourney (right column) produces images that are quite distinct from the reference image, but that resemble other images of the townscape of Djenné, showing markets, houses, and people in transit. This set of images shows the compositional nature of Midjourney's generated images: in each case some version of the mosque is recognisable, but in the background, shot in shadow and occasionally at oblique angles. People in the foreground feature in a quasi-cinematic way: in two cases, one or two people appear close to the presumed camera, as though on a

Figure 2. Old Towns of Djenné (Mali), date: 18/02/2005, author: Francesco Bandarin, copyright: © UNESCO CC3 license.

journey, while more distant figures appear as accidental subjects. In all cases, people appear in some variant of an assumed local dress.

For the 'M-M' reading of the source image, we obtained the following:

> *photograph of a large sand castle with people walking in front of it*, **Building center, Sky, Cloud, Travel, Landscape, Sand, Aeolian landform, Facade, History, Ancient history, Archaeological site, Historic site, Art, Arch., Soil, Horizon, Singing sand, Tourism, Castle, Desert, Tourist attraction. Colors: #bb9667, #b7c1c3, #b78e5c, #9b7344, #977649, #6f4b1f, #aa9373, #694e26, #634d2d, #8d795a** Shot with a E3700, at a resolution of 300 pixels per inch, year 2005, exposure time of 5/1806, Flash did not fire, auto mode, focal length of 27/5

The first part, in italics, represents the BLIP2 caption; the second, in bold, a textual representation of properties extracted from the Google analysis; and the third, metadata properties of the source image. The misrecognition of the mosque as a sandcastle in the machinic reading can be attributed to an alignment of the language of castles with similar visual patterns in the training data. Figure 4 represents the outputs, following the same pattern as Figure 3.

In each case, the anchoring characteristic is the first part of the prompt, 'large sandcastle'. In one case (Midjourney, bottom), there are recognisable aspects of the source image, including the mosque's exterior ornamentation. But in most cases the 'castle' produced more closely resembles a Disneyfied castle caricature, diverting quite starkly from the rectilinear form of the mosque. A recurring similarity in most of the images produced is a lack of surrounding built context: both the mosque and the castle appear to be isolated monumental objects in the frame.

Figure 3. Machinic image outputs using the UNESCO World Heritage Site description (models used: SDXL, Realistic Vision, Midjourney v5).

In other respects, and despite the prompt specifying colours, camera type, and exposure time, the reference to a sandcastle appears to over-determine the colour palette and saturation level. Compared to Figure 3 (H-M), in Figure 4 (M-M) the sand, of both castle and foreground, is lighter, and the sky clear rather than hazy. Keywords such as 'history' and 'archaeology' also change the sense of scale and context, with the implied camera position being now more distant. The scene is also deracinated: the form of the 'castle' is drawn from a wide range of typological and stylistic references, and though diminutive, the 'people' referenced in the prompt are dressed in global rather than 'local' attire, tourists who apprehend the monumental structure rather than locals who live around it. The presumed holder of the gaze is, in other words, no longer solely a figure imagined as behind the camera, but firmly embedded within it.

Figure 4. Machinic image outputs using a machinic prompt (prompt generation uses BLIP2, Google Vision API, and metadata, and image generation uses SDXL, Realistic Vision, and Midjourney v5).

Old City of Sana'a

Figure 5 shows the original UNESCO World Heritage Site image (top) of the Old City of Sana'a, along with two generated outputs from Midjourney 5.2: the first (middle, human-machine or 'H-M') is the result of the UNESCO, human-authored description used as a prompt, and the second (bottom, machine-machine or 'M-M') the output of the machine-generated prompt. In this case, the UNESCO description places emphasis on the city-scape, with phrases like 'rammed earth and burnt brick towers' and 'densely packed houses, mosques', but also specifies colours—'white gypsum', 'bistre colored earth', 'green bustans'. The BLIP generated prompt correctly identifies the image subject—'old city of Yemen'—but also the frame: 'an aerial view.'

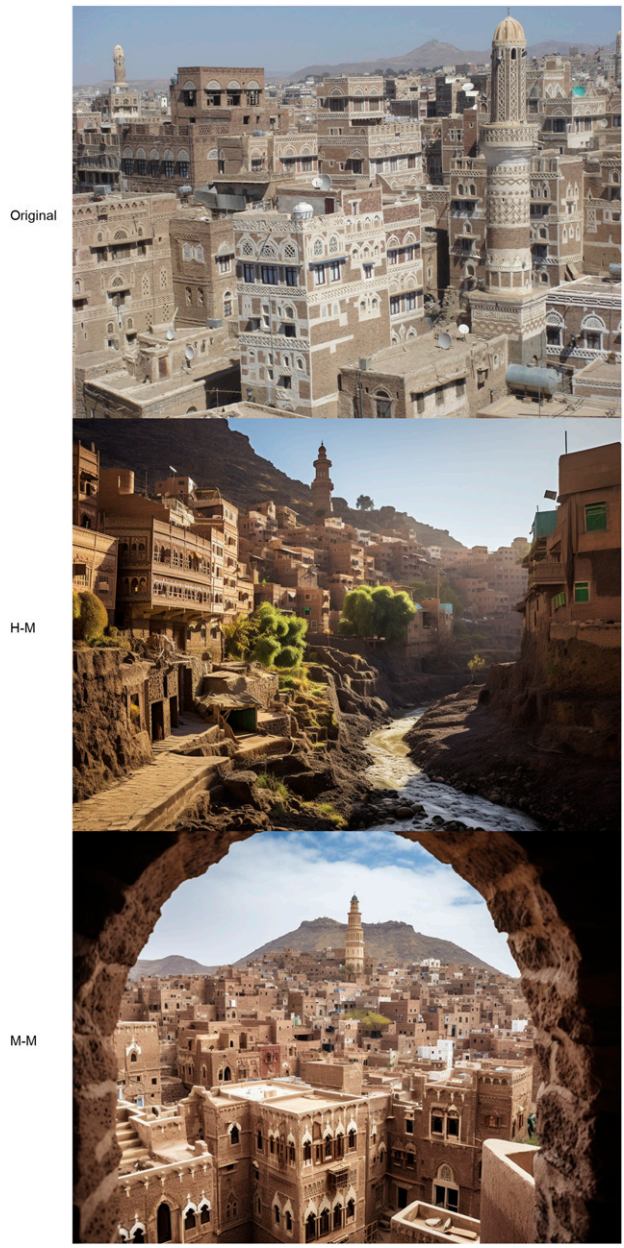

Figure 5. Top: Old City of Sana'a (Yemen), author: Maria Gropa, copyright: © UNESCO, reproduced with a CC3 license. Middle: A synthesised image using the UNESCO World Heritage Site description as prompt (model: Midjourney). Bottom: A synthesised image using a machinic prompt (prompt uses BLIP2, Google Vision API, and metadata, and image generation uses Midjourney v5).

Here the machine-generated prompt was:

> photograph of an aerial view of the old city of yemen, **Building center, Sky, Daytime, Window, Architecture, Landscape, City, Urban design, Landmark, Cityscape, Facade, Roof, Human settlement, Urban area, Medieval architecture, Metropolis, Arch, Mixed-use, Archaeological site, Ancient history, Historic site, History, Turret, Dome, Town, Monument, Bird's-eye view, Tourism, Classical architecture, Holy places. Colors: #cdc3b8, #cfc2b1, #ab9b8a, #a69b93, #e7ddd2, #83766e, #887868, #e9dccb, #5f534d, #3f342e.** *Shot with a DSC-T9, at a resolution of 72 pixels per inch, year 2009, exposure time of 1/500, no flash, focal length of 1139/100*

As with the Midjourney outputs for Djenné, both generated outputs show a tendency to emphasise geographic features identified in the textual description or source image. Mountains are exaggerated, and in the 'H-M' case parts of the city hug a cliff-face and overlook a river, in sharp contrast to the source photo. Tonally, the 'H-M' image also employs stronger use of contrast (brightly illuminated buildings on the left compared to those in shadow on the right), and a greater colour dynamic—browns, vivid blues, and varying greens—reflects the especially chromatic verbal description ('spacious green bustans').

The 'M-M' image, on the other hand, is strikingly similar, both in broad elements of architectural form and image composition, to the source. Just as with 'giant sandcastle' in the case of Djenné, here both the identification of aspect ('aerial view') and location ('old city of yemen') work to determine scale, perspective, and chromatism of images for all three models. In the case of the selected Midjourney image, the identification of an 'arch' object by the Google API—barely discernible in the source image—is brought into the fore as a photographic conceit, a 'found' frame for the distant cityscape. Though not evident in this source image, even another of the UNESCO images of Sana'a employs the same framing device—a convention of the 'serious' or expert photographer the machine has learned to reproduce. Despite the inclusion of a palette extracted from the source, though, the colours of the sky and buildings are once again more lurid and saturated than those that appear in the official 'expert' gaze—a kind of machinic equivalent to an Instagram filter designed to appeal instead to some imagined, would-be tourist to the city.

This last feature is unsurprising for several reasons: the training sets include more 'tourist' than 'expert' images, reinforced by the very inclusion of the term 'tourism' alongside 'archaeological history' in the generated prompt; more contemporary images featured in those training sets also use a greater colour range than even those from the 2000s decade; the reference to a specific location; and finally, Midjourney itself is a commercial system that has been 'fine-tuned' to produce arresting images precisely through use of high contrast. And yet, in the final case we discuss here, this effect is in fact reversed.

Tombs of Buganda Kings at Kasubi

Figure 6, featuring representations of the Tombs of Buganda Kings at Kasubi, uses the same pattern as Figure 5: at the top is the original UNESCO World Heritage Site image, followed by two synthetic images, this time generated by Stable Diffusion XL, selected

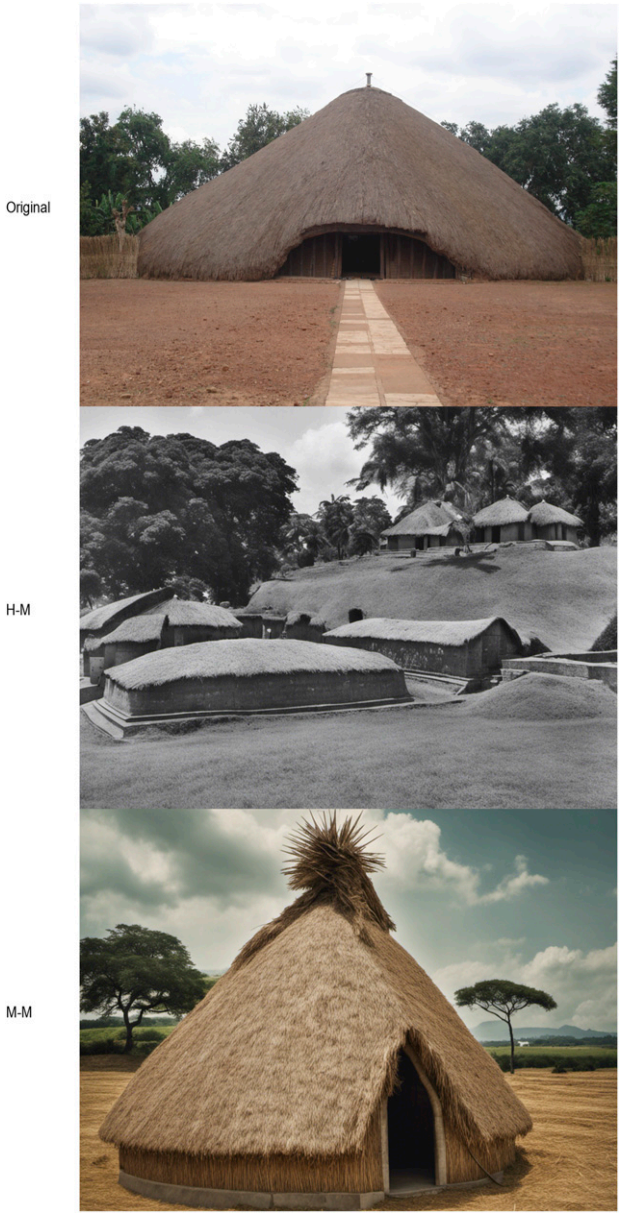

Figure 6. Top: Tombs of Buganda Kings at Kasubi (Uganda), author: Lazare Eloundou Assomo, copyright: © UNESCO, reproduced with a CC3 license. Middle: A synthesised image using the UNESCO World Heritage Site description as prompt (model: SDXL). Bottom: A synthesised image using a machinic prompt (prompt uses BLIP2, Google Vision API, and metadata, and image generation uses SDXL).

for the purpose of contrast. The middle image (H-M) is again produced from the UNESCO textual prompt, while the bottom image (M-M) is from a prompt constructed from machine-generated captions and image metadata. The UNESCO description in this case emphasises the materiality of structures, with the phrases 'organic materials' and 'wood, thatch, reed, wattle, and daub', but also references form: 'circular and surmounted by a dome'. The machinic prompt locates the structure and identifies the image as a 'photograph of the roof (is) made of straw'.

Machine-generated prompt:

> photograph of the roof is made of straw, **Building center, Cloud, Sky, Land lot, Tree, Thatching, Shade, Grass, Tints and shades, Roof, Monument, Triangle, Soil, Historic site, Symmetry, Landscape, Building material, Hut, House. Colors: #83726b, #9a7360, #392d29, #d7b9a4, #f2f3f6, #7c685d, #211918, #bb9783, #645650, #6b584b.** *Shot with a DSC-W50, at a resolution of 72 pixels per inch, year 2007, exposure time of 1/80, Flash did not fire, auto mode, focal length of 47/5*

The first photograph of the Kasubi tombs, representing a front elevational aspect to the main structure, focuses primarily on the structure's symmetry, materiality ('thatch and reed' in particular), and form, while the tight framing of the camera angle and the relative absence of other objects and context add a sense of scale, creating a sense of monumentality in the fairly austere building. The photograph of the single structure devoid of context emphasises a monumentality that is not reflected in the UNESCO description, which instead identifies intangible aspects of the tomb, including the continuity of its use and its associated meaning. These non-visual cues acknowledge that the building's aesthetics and form are not solely constitutive of its value as a heritage site.

One of the generated images was a black and white photograph, which we speculate is in response to the specific mention of dates (1882/1884) in the prompt potentially directing the colour scheme. The tonality, frame, and context of the H-M image are closest to photographs of late 19th- and early 20th-century archaeological surveys.

The subject of the M-M image is notionally closer to the original in terms of morphology, materials, and a focus on roof form. The foreground landscape echoes the materiality of the subject, while the background reproduces vegetation and tonality often depicted in images of the African savanna. The tight framing of the structure in the photograph and the difficulty in assigning a sense of scale mimic the original image, but the central difference between the two is in framing the subject, which shifts the emphasis from the monumental in the original to something more vernacular in the M-M image.

Heritage and the machinic gaze

The algorithmic reading and synthesis of the three UNESCO World Heritage Sites offers an interesting counterpoint to UNESCO's own textual descriptions. All three site images, when read via BLIP-2, focused on the descriptions of form, scale, material, and composition, erasing any sense of aesthetic judgement or valuation and instead generating descriptions for precision and conciseness with varying levels of accuracy. For instance,

while the caption generated for the historic centre of Sana'a accurately identified 'an aerial view of the old city of Yemen', the caption generated for the photograph of the Old Towns of Djenné was 'a large sand castle with people walking in front of it', while the photograph of the Tombs of Buganda Kings at Kasubi was 'the roof is made out of straw'. The misrecognition of the Great Mosque of Djenné as a sandcastle reflects perhaps most clearly the distortion introduced by a machinic reading of this kind. However, even the simplification of the Kasubi tombs to essentially an image of a roof allows us to reflect upon our own interpretation of the images as sites of globally recognised heritage. The second layer of algorithmic reading of the image, via Google Vision's API, followed a mathematical extraction based on probabilistic interpretation. In each of the images, elements such as 'sky', 'grass', and 'building center' were identified, alongside other identifying descriptors such as 'medieval architecture', 'arch', and 'archaeological site', but also specific descriptions such as 'Classical architecture' or 'Byzantine architecture'. Occasionally seemingly contradictory descriptors would be generated for the same image, once again illustrating the slippage between image and text in the absence of a referent informing the machinic gaze.

The anamorphic properties of the machinic gaze play out in all three cases, but especially so with Djenné. The mosque is interpreted as a 'giant sand castle', and subsequent image synthesis then renders this as an artefact that substantially deviates from the original object. However, the machine ignores cues in the image, focussing on the form of the dominant subject, and inferring the likely class of building based on a reading of pixels and profiles: castle rather than mosque. This is due to the way images are processed iteratively: first with coarse filters that aim to identify, for example, horizontal and vertical lines, then with finer filters that progressively distinguish more subtle gradations in form and colour. Buildings—as relatively geometrically regular objects of a certain scale—are likely to be seen as alike, regardless of functional distinctions between, for example, a place of worship and a playful structure designed to imitate a castle. Such distinctions, if they feature at all, depend in turn upon the relative mass of images and labels in the training set. Hence the apparent confusion between a certain type of mosque and a sandcastle reflects the proportionate mass of labelled images of Djenné, relative to other sites—and the corresponding value attributed, in the human (tourist and archaeological) gaze, to that site. To 'correct' this error would involve different practices of touristic attention (or modified weightings of the training data) to better 'align' this vision with human expectation. Conversely, it is precisely this orthogonal or anamorphic perspective that in turn reflects upon existing practices of human observation and perspective—the privileging of certain sites over others, the concentration of canonical representations of 'mosques' and 'castles', and the re-projection of localised settings into the global imaginary of tourism and heritage.

In the context of the heritage image, how should we describe the process by which, for instance, the old towns of Djenné become instead whimsical sandcastles that impossibly dwarf human characters in the foreground? No existing heritage taxonomic overlay can quite work to make sense of these creations, and even existing artistic nomenclature would struggle to 'locate' these examples of machinic heteroscopia. This step of algorithmic reading, which is devoid of the human 'expert' or the 'tourist' gaze that relies on

a constant referent to ideas of heritage value derived from architectural and/or archaeological aesthetics and classifications, but which instead focuses purely on pixels of an image, reveals the extent of meaning we implicitly attach to images of heritage sites. Deploying the machinic gaze towards photographs allows us to occupy a position of tourist or expert—or in some cases both—but in each case, we can reflect upon the presumed author/generator of the image. On the other hand, multiple historic and visual referents are embedded within each of the five UNESCO site descriptions. Read alongside the description, the image of Djenné is inscribed with multiple aesthetic judgements and associated ideas of heritage value.

The privileging of the visuality and aesthetics of heritage sites in UNESCO is, we argue, distorted and refracted through the machinic gaze, and through the operations we identify as anamorphosis and heteroscopia. In highlighting elements of both similarity and difference, through visual representation, the fetishisation of architectural form, ornamentation, and material can be examined through both sets of images. In the first set, where the human textual description is used as visual description/reinscription, we observe a greater degree of diversity in both subject and framing, but consistent in the images produced is a privileging of a certain kind of aesthetic that aligns to the idea of heritage value being inscribed and prescribed visually and materially. In the second set, which is produced through a machinic reading and resynthesis, even though the subject of the image shifts substantially, the framing does not.

We argue that heteroscopia and anamorphosis help to cluster and aggregate these features into refracted and concentrated delineations that otherwise exist as more diffused tendencies or proclivities: how the tourist and the expert see. These tendencies appear more or less evident across two of the site/model/prompt combinations. Sana'a (with Midjourney) is reproduced through something like the tourist gaze—imagined at a distance, with saturated colour—while outputs prompted by Kasubi (with SDXL) prompts appear closer to an expert's view—muted palette, with the photographic subject brought to the foreground. The Djenné images share elements of both, but veer into alternative registers of the cinematic and fantastical. In calling for such interpretations, the machine here acts to bring these gazes themselves into focus. And with the act of interpretation itself, we move invariably away from attention to purely quantitative variances—inherent in the very mechanisms by which machine learning techniques aim to approximate a training set—to emphasise instead a process of human judgement and critique.

We conclude on a speculative note about the effects of this process. In Lacan's treatment of the gaze, which MacCannell (2001) draws upon, its significance is the drawing back in of the viewing subject into the picture or tableau (Lacan, 1998). It is the subject who, alongside the image under apprehension, at a critical moment perceives themselves as being also observed, as an object that appears in the eyes of others. The emergence of computer vision, machine learning, and generative AI exacerbates this reflexive moment. The human gaze—especially in its tourist or heritage genres—becomes aware of itself in its particularity, as a thing both distinctive and available in turn as object for consumption by other viewers. The combined operation of heteroscopia and anamorphosis here performs a kind of double act with respect to human ways of seeing.

Firstly, it points to the impossibility of any privileged and original specular *morphosis*, or authoritative gaze. Secondly and conversely, it points to the possibility of any given form of gaze becoming itself treated as quotable reference material, and in that process, also becoming objectified.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Vanicka Arora https://orcid.org/0000-0001-8733-4510

Supplemental Material

Supplemental material for this article is available online.

Note

1. To aid reproduction, and to support similar studies, the scripts and datasets have been published on GitHub: https://github.com/liammagee/reframing-built-heritage-through-the-machinic-gaze

References

Azar M, Cox G and Impett L (2021) Introduction: ways of machine seeing. *AI & Society* 36: 1093–1104.

Barauah A (2017) *"Travel imagery in the age of Instagram: An ethnography of travel influences and the "online tourist gaze"*. Master's Thesis, SOAS, University of London, UK.

Beck AS and Sørensen TF (2017) 32 eyes: archaeology and the subjective gaze. *Journal of Contemporary Archaeology* 4(1): 131–152.

Benjamin W (1986) *Illuminations*. New York: Random House.

Bohrer FN (2011) *Photography and Archaeology*. London: Reaktion Books.

Chadha A (2002) Visions of discipline: Sir Mortimer Wheeler and the archaeological method in India (1944-1948). *Journal of Social Archaeology* 2(3): 378–401.

Champion E (ed) (2019) *The Phenomenology of Real and Virtual Places*. New York and London: Routledge.

Crary J (1990) *Techniques of the Observer*. Cambridge, MA: MIT Press.

Croitoru F-A, Hondru V, Ionescu RT, et al. (2023) Diffusion models in vision: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45(9): 10850–10869.

Denicolai L (2021) The robotical gaze: a hypothesis of visual production using technological image/thinking. MediArXiv preprint. Available at: https://mediarxiv.org/undh5/ (accessed 19 February 2024).

Dicks B (2000) Encoding and decoding the people: circuits of communication at a local heritage museum. *European Journal of Communication* 15(1): 61–78.

Forte M (2007) Ecological cybernetics, virtual reality, and virtual heritage. In: Cameron F and Kenderdine S (eds) *Theorizing Digital Cultural Heritage: A Critical Discourse*. Cambridge, MA: MIT Press, 389–407.

Hacıgüzeller P (2012) GIS, critique, representation and beyond. *Journal of Social Archaeology* 12(2): 245–263.

Jaireth S (2000) To see and be seen: the heteroscopia of Hindi film posters. *Continuum* 14(2): 201–214.

Kittler F (2010) *Optical Media*. London: Polity.

Lacan J (1998) *Book XI: The Four Fundamental Concepts of Psychoanalysis. Trans. A Sheridan*. New York: WW Norton & Company.

Li J, Li D, Savarese S, et al. (2023) BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint. Available at: https://arxiv.org/abs/2301.12597 (accessed 19 February 2024).

MacCannell D (2001) Tourist agency. *Tourist Studies* 1(1): 23–37.

Midjourney (2022) We're officially moving to open-beta. https://discord.gg/midjourney, **Please read our directions carefully** or check out our detailed how-to guides here: https://midjourney.gitbook.io/docs, https://twitter.com/midjourney/status/1547108864788553729 (accessed 5 October 2023).

Moshenska G (2013) The archaeological gaze. In: González-Ruibal A (ed). *Reclaiming Archaeology: Beyond the Tropes of Modernity*. London: Routledge, 211–219.

Mostaque E (2022) Stable diffusion public release. Available at: https://stability.ai/blog/stable-diffusion-public-release (accessed 10 April 2022).

Mulvey L (2013) Visual pleasure and narrative cinema. In: Penley C (ed). *Feminism and Film Theory*. London: Routledge, 57–68.

Niklasson E (2017) The Janus-face of European heritage: revisiting the rhetoric of Europe-making in EU cultural politics. *Journal of Social Archaeology* 17(2): 138–162.

Offert F and Phan T (2022) A sign that spells: DALL-E 2, invisual images and the racial politics of feature space. arXiv preprint. Available at: https://arxiv.org/abs/2211.06323 (accessed 19 February 2024).

Oh Y (2022) Insta-gaze: aesthetic representation and contested transformation of Woljeong, South Korea. *Tourism Geographies* 24(6–7): 1040–1060.

Parisi L (2019) The alien subject of AI. *Subjectivity* 12: 27–48.

Ramesh A, Dhariwal P, Nichol A, et al. (2022) Hierarchical text-conditional image generation with CLIP latents. arXiv preprint. Available at: https://arxiv.org/abs/2204.06125 (accessed 19 February 2024).

Salvaggio E (2022) How to read an AI image. *Cybernetic Forests*, 2 October. Available at: https://cyberneticforests.substack.com/p/how-to-read-an-ai-image (accessed 10 July 2023).

Santos Paula Mota (2016) Crossed gazes over an old city: photography and the 'Experientiation' of a heritage place. *International Journal of Heritage Studies* 22(2):131–144 doi: 10.1080/13527258.2015.1108925

Schuhmann C, Beaumont R, Vencu R, et al. (2022) LAION-5B: an open large-scale dataset for training next generation image-text models. arXiv preprint. Available at: https://arxiv.org/abs/2210.08402 (accessed 19 February 2024).

Smith L (2006) *Uses of Heritage*. London: Routledge.

Sterling C (2016) Mundane myths: heritage and the politics of the photographic cliché. *Public Archaeology* 15(2–3): 87–112.

Sterling C (2019) *Heritage, Photography, and the Affective Past*. London: Routledge.

UNESCO (2008) The criteria for selection. Available at: https://whc.unesco.org/en/criteria/ (accessed 5 October 2023).

UNESCO (2023) World Heritage List. Available at: https://whc.unesco.org/en/list/ (accessed 5 October 2023).

Urry John (1990) The 'consumption' of tourism. *Sociology* 24(2): 23–35. https://doi.org/10.1177/0038038590024001004

Urry J and Larsen J (2011) *The Tourist Gaze 3.0*. London: Sage.

Waterton E (2009) Sights of sites: picturing heritage, power and exclusion. *Journal of Heritage Tourism* 4(1): 37–56.

Watson S and Waterton E (eds) (2016) *Culture, Heritage and Representation: Perspectives on Visuality and the Past*. London: Routledge.

Wickstead H (2009) The uber archaeologist: art, GIS and the male gaze revisited. *Journal of Social Archaeology* 9(2): 249–271.

Winter T (2006) Ruining the dream? The challenge of tourism at Angkor, Cambodia. In: Meethan K, Anderson A and Miles S (eds). *Tourism Consumption and Representation: Narratives of Place and Self*. Wallingford: CABI, 46–66.

Zhang J, Huang J, Jin S, et al. (2023) Vision-language models for vision tasks: a survey. arXiv preprint. Available at: https://arxiv.org/abs/2304.00685 (accessed 19 February 2024).

Author biographies

Vanicka Arora is Lecturer in Heritage at University of Stirling. She looks at the intersections of heritage, disasters, urbanisation, and globalisation, with an empirical focus on South Asia. She has recently turned her attention towards the methodological possibilities of generative AI in visual studies and heritage.

Liam Magee is Associate Professor at the School of Humanities, Communications & Arts, Western Sydney University. He works on topics of automation, digital media, urban sustainability, and social inclusion, and recently has focussed on the relationship between generative AI and subjectivity, aesthetics, and labour.

Luke Munn is a Research Fellow in Digital Cultures & Societies at the University of Queensland. His wide-ranging work investigating digital cultures has been published in more than 40 articles in highly regarded journals and in six books, including most recently *Automation Is a Myth* (2022), *Red Pilled* (2023), and *Technical Territories* (2023).