From Galleries to Generators: Applying Museum Empathetic Strategies to Reduce Confirmation Bias in Chatbots

Cassandra Kist[1]

[1]University of Strathclyde, Computer and Information Sciences

Abstract

Museums have long positioned themselves as socially responsible institutions, designing heritage interpretation to foster inclusion, and cultivate cross cultural understanding. The 'affective-turn' in museum practice has expanded heritage interpretation beyond text toward multimodal, embodied, sensory, and emotive experiences that support visitors' reflective meaning making. In this provocation, I argue that general-use Generative AI chatbots (such as ChatGPT and Co-pilot) should take inspiration from museum interpretive techniques to address growing risks of confirmation bias arising from biased queries, belief consistent personalisation, and overly agreeable chatbot response styles. Drawing on museum strategies for cultivating empathy – encompassing perspectivity, reflective action, and sensory, and affective connection – I outline how these interpretive techniques could inform chatbot interaction design. I conclude with design directions for perspective expanding chatbots that mitigate bias not only through cognition, but also by privileging users' meaning making as an emotive process.

Keywords

Museum interpretation, Empathy, Confirmation bias, Chatbots

1. Introduction

In the heritage sector there has been a drive to inform and even perpetuate responsible practices in the development and implementation of AI technologies [1]. This stems from the social value-driven nature of museums to be in the service of society, upholding ideals such as inclusion, cross-cultural understanding, and environmental sustainability. Recently, researchers in Human Computer Interaction (HCI) have suggested that ways of working e.g., in participatory ways with communities can be brought into AI-design and development processes [2]. In this provocation paper, I argue that by designing general use generative AI chatbots less like search engines and more like museum interpreters we can create novel solutions to overcome confirmation biases. I theorise that interpretation frameworks for cultivating empathy in museums related to perspectivity, reflective action, and sensory, and affective connection, could inform the interaction design of chatbots to widen user perspectives. First, I share a museum-based perspective on empathy and related interpretive techniques; 2) then, I outline key challenges caused by chatbot architecture relating to confirmation bias; and 3) I critically reflect on how these museum interpretation techniques could inform interaction designs of generative AI chatbots.

In museums, empathy is often believed to support the identity building practices of visitors and as potentially leading to pro-social behaviours [3]. As a result of the 'affective turn' in museology, museums have realised the importance of interpretation that extends beyond an emphasis on cognition to also value visitors' meaning making practices through emotions such as empathy [4]. As part of this shift, museum interpretation techniques (ways of supporting users to engage and learn through heritage) have shifted away from text-based interpretation to more immersive, embodied, and sensory forms of communication to support emotional connections and empathetic experiences [4].

EmpathiCH'26 Workshop Co-located with CHI'26 Conference on Human Factors in Computing Systems, April 13–17, 2026, Barcelona, Spain

cassandra.kist@strath.ac.uk (C. Kist)

https:https://pureportal.strath.ac.uk/en/persons/cassandra-kist/ (C. Kist)

0000-0001-9960-2236 (C. Kist)

2. Defining and Cultivating Empathy in Museums

Museums have long held an educational mission to inform visitor perspectives and widen understanding of complex and difficult topics that may be contentious and sit uncomfortably with visitors' positive sense of identity [5]. Difficult heritage, however, are labelled as 'difficult' or 'challenging' due to not only its subject matter (often connected to conflict, trauma), but also how it challenges visitors to consider what this 'knowledge does to us—or what we do with it' [6, p. 7]. As part of interpreting difficult heritage, cultivating empathy has been a central concern and an idealised way to support critical learning. Empathy has been described in various ways, but commonly in the heritage sector empathy is described as 'feeling for another' while also encompassing a critical, reflective understanding of what it means to live in relation to someone else's experiences [7]. This lens positions empathy not as a one-to-one perspective taking but a form of engagement that balances both emotional connection and critical cognitive reflection. The balancing of emotion and cognition aligns with the concept of 'higher-level empathy' [8], or what others researchers have similarly coined as an 'ethical witnessing' of- [7] or an 'intimate encounter' [9] with another person's experiences. Comparatively, this is differentiated from a 'lower level empathy', described as an unconscious reaction that mimics another person's emotional state without prompting a critical understanding or reflection [8].

Professionals and heritage-researchers have theorised that to support higher-level empathy, designing for immersive engagement with personal narratives is an essential technique for promoting visitors' emotional and critical connections. For example, immersive engagement with personal narratives has been cultivated by enabling visitors to 'meet' and interact with past characters, act as past characters, and create embodied experiences that enable encounters with first-person perspectives and lived experiences. Researchers have captured the success of such initiatives when investigating the performance of historic people in museum contexts and its effects on visitor learning and empathy: 'Empathy was evidenced as a powerful emotional tool giving depth of insight – into the lives of individuals especially – that was difficult to achieve through other more formal, cognition-based modes of learning' [10, p. 115]. Other examples of facilitating emotional connections to personal narratives are through video-games that enable role-play and perspective shifting, bringing different perspectives and experiences closer to players [11].

However, as further explained by the researchers investigating heritage performance and empathy, 'empathy can, in certain circumstances, offer a rather partial 'monocular' reading of events, narrowing our vision…' [10, p. 115]. Multi-perspectivity here, and as suggested by other researchers/practitioners, is essential to avoid mono-lithic perspectives - by having multiple voices, experiences and perspectives represented, enabling a zooming-out from a singular perspective [12]. Museum interpreters also call attention to the poetics of the exhibition and interpretation space in terms of how these elements support emotional connections and critical reflection of visitors: for instance the use of serialisation (scattering multiple mini-narratives throughout an exhibit) has been theorised as supporting engagement with a range of perspectives and narratives while emphasising the fragmentary nature of heritage interpretation [13]. Other professionals have highlighted the importance of interactive prompts to cultivate reflections. For instance, this has been enabled through direct invitations for visitors to stop, reflect and share, for example through sticky notes, visitor cards, and digital interactives. On the other hand, critical reflection can also be encouraged by sensory experiences such as silences and absences: For example in the digital heritage project - 'With New Eyes I See' (WNEIS), multimodal, itinerant outdoor projections were used to foreground embodiment, empathy, and silence as visitors moved through layered narratives of WWI histories [14]. This approach embraced ambiguity and allowed meaning to emerge through sensory and affective engagement, prompting visitors' critical reflections.

3. Chatbots and Confirmation Bias

In the design of chatbots, confirmation bias has been an ongoing issue at the intersection of social practice, information behavior, and interface and algorithmic design. As there is increasing reliance on

generative AI agents in everyday information seeking and in higher-stakes context (such as healthcare, law etc.) working towards solutions for confirmation bias is urgent. Confirmation bias refers to users' preferences for information that agrees with existing perspectives and ideas as opposed to those that refute or challenge them, which is a practice rooted in psychology. This behavior is partially due to the fact that information which aligns with one's existing ideas and worldview can be comforting while information that challenges existing ideas can create unease [15]. According to cognitive psychology, confirmation bias behaviors are associated with strategies to minimise mental effort [15]. However, such behaviours (seeking confirmation information and avoiding contradictory information) as suggested by [16] can be shaped by individual factors (such as certain beliefs e.g., Anti-vaccine and confidence in one's own expertise) [17], and situationally, by how information tasks are framed [18].

Crucially, technologies themselves contribute to the amplification or mitigation of confirmation bias. The affordances of chatbots - especially generative AI do not merely reflect confirmation biases but can intensify them through technical architecture, personalisation strategies, and interaction design [15, 16, 19]. In the following, I synthesise how key architectural and interactive features of generative AI systems sustain or exacerbate confirmation bias, and I summarise emerging mitigation techniques. The first challenge is training data and model architecture. Large language models are trained on massive, heterogeneous data that embed historical and cultural biases. When prompted with a belief-laden or leading question, models tend to reproduce those entrenched patterns [15]. As a result, these can appear to confirm user assumptions, strengthening the belief in their perspective [20]. To mitigate this, researchers suggest that there should be retrieval methods which surface multiple perspectives and make transparent where information comes from [15]. To do so, would require a robust training data that reflect perspectives which may be on the margins, and an element of transparency when they are absent [15].

A second major challenge is the hyper-personalisation of chatbot interactions - hyper-personalisation makes chatbots increasingly tailored to users' pre-existing preferences, amplifying confirmation bias [20, 16]. Personalisation becomes a form of 'echoing' reinforcing worldview-consistent information while suppressing alternatives [20, 16]. Researchers recommend using mitigation techniques proposed for other online contexts such as self-nudges that encourage users to explore belief-inconsistent viewpoints [16]. Comparatively, [15] suggests optional 'balanced response' that automatically provide alternative arguments. Beyond these technical suggestions, supporting AI-literacy by helping users formulate unbiased queries has been suggested as mitigating biases related to hyper-personalisation [16].

Finally, a third challenge relates to interface design and 'sycophancy': generative AI models often produce overly agreeable or user-affirming responses known as sycophancy, as a result of the system being designed to maximise user satisfaction [19]. Users interpret agreement as validation, deepening confirmation bias and narrowing their potential to engage with contradictory information. Others have suggested using consider-the-opposite techniques and cultivating users' AI literacies (through exposure to misinformation tactics), using persuasive evidence-based counter-arguments [16] and fine-tuning against over-agreement patterns to mitigate this challenge [15].

4. Informing Interaction Design with Chatbots

Current approaches to mitigating confirmation bias in chatbots tend to draw primarily from cognitive psychology, information behavior, and computational interventions. Yet these framings largely treat users as abstract information processors rather than embodied, situated individuals whose beliefs are shaped through narrative, emotion, and social context. In contrast, museum interpretation has long grappled with the challenge of helping visitors encounter perspectives beyond their own, often across vast temporal, cultural, or experiential distances. These interpretive traditions offer rich, underutilised design strategies that may provoke new ways of imagining how chatbots could address confirmation bias, including individual narratives, balancing multi-perspectivity, and provoking reflection through sensory experiences. Museums regularly use first-person/personal narratives - sometimes facilitated through embodiment/performance, vignettes, and micro-stories to evoke empathy and open visitors to

alternative viewpoints (e.g. [10]). Ethnographic vignettes, for example, are designed not to 'teach facts' but to invite emotional proximity to another person's lived reality. These techniques often operate by foregrounding subjective experience rather than abstract arguments, inviting the visitor into an event, and showing how meaning is shaped by one's social world, body, and environment. Translating this into chatbot design suggests a shift from purely factual responses ('consider-the-opposite' [16]) toward emotionally grounded counter-narratives that could gently reframe the user's viewpoint. Rather than opposing a user's belief directly, a chatbot might adopt a practice similar to museum storytelling: offering a vignette - a lived example that expands understanding using affective connections.

Another tactic to challenge confirmation bias is designing for multi-perspectivity: Museums often aim to design for balanced perspectivity by placing diverse voices and plural interpretations side by side, without collapsing differences or claiming neutrality [12]. Crucially, they do this with attention to emotion - visitors are guided to care about perspectives beyond their own. Designing chatbots to counter confirmation bias might similarly require presenting multiple perspectives and ideas not as a list of facts but as described above - lived, situated viewpoints. This would cultivate affective bridges that help users appreciate why someone might hold a different view, incorporating both affective and cognitive understandings. However, as heritage interpretation reminds us, pluralizing perspectives requires the right data, the right ethical commitments, and careful attention to power [21].

Finally, heritage interpretation also uses the material environment, objects, and even silences to shape and encourage visitors' critical reflections and emotional connections. These sensory and spatial strategies acknowledge visitors as embodied and affectively situated. While chatbots cannot reproduce physical space, they might use similar techniques such as deliberate pauses and encouraging reflective prompts to slow cognitive momentum and enable critical thinking. Invoking 'silence' in chatbot interactions could also translate into purposeful withholding - moments where the agent does not rush to assert certainty but opens an affective space for the user to reflect, question or re-examine an assumption.

Emerging heritage conversational agents - such as virtual veterans, talking specimens, or embodied objects - serve as prototypes for how narrative, persona, and emotional engagement can be designed for with the potential to reshape visitor understanding. Though not designed for mitigating confirmation bias, they show the potential of perspective-taking through personas/lived-experiences, emotional resonance through narrative, and guidance through gently structured, reflective conversational flows. These systems demonstrate (to different extents) that affective-driven design can open visitors to unfamiliar ideas by balancing emotional connection and cognition needed to counter confirmation bias.

5. Conclusion

To conclude, in this provocation paper I suggest that mitigating confirmation bias may require us to design chatbots less like search engines and more like museum interpreters. Not necessarily to 'fix' users' beliefs, but to widen their perspective and encourage consideration for alternative viewpoints and ideas. Heritage interpretation offers inspiration for how to do this ethically, intentionally, and affectively: Specifically, interpretive techniques for encouraging empathy through perspectivity, reflective action, and affective connections could support overcoming confirmation bias by honoring users' meaning making as not only a cognitive but also an emotional process. As chatbots increasingly mediate how people understand themselves and their world, these interpretive strategies may become essential tools for responsible AI design.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] UNESCO, Recommendation on the ethics of artificial intelligence, 2021. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455.

[2] H.-Y. Huang, C. C. S. Liem, Social inclusion in curated contexts: Insights from museum practices, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, p. 300–309. URL: http://arxiv.org/abs/2205.05192. doi:10.1145/3531146.3533095, arXiv:2205.05192 [cs].

[3] J. Kidd, J. Sayner, Intersections of silence and empathy in heritage practice, International Journal of Heritage Studies 25 (2019) 1–4. URL: https://doi.org/10.1080/13527258.2018.1475414. doi:10.1080/13527258.2018.1475414.

[4] M. Varutti, The affective turn in museums and the rise of affective curatorship, Museum Management and Curatorship 38 (2023) 61–75. URL: https://doi.org/10.1080/09647775.2022.2132993. doi:10.1080/09647775.2022.2132993.

[5] S. Macdonald, Difficult Heritage: Negotiating the Nazi Past in Nuremberg and Beyond, Routledge, 2010. Google-Books-ID: Z199AgAAQBAJ.

[6] E. Lehrer, C. E. Milton, Introduction: Witnesses to Witnessing, Palgrave Macmillan UK, London, 2011, p. 1–19. URL: https://doi.org/10.1057/9780230319554_1. doi:10.1057/9780230319554_1.

[7] E. A. Kaplan, Empathy and Trauma Culture: Imaging Catastrophe, Oxford University Press, 2011, p. 255–276. doi:10.1093/acprof:oso/9780199539956.003.0016.

[8] A. Coplan, P. Goldie, Introduction to Empathy: Philosophical and Psychological Perspectives, Oxford University Press, 2011. doi:10.1093/acprof:oso/9780199539956.001.0001.

[9] J. Bonnell, R. I. Simon, "difficult" exhibitions and intimate encounters, Museum and Society 5 (2007) 65–85. URL: https://journals.le.ac.uk/ojs1/index.php/mas/article/view/97. doi:10.29311/mas.v5i2.97.

[10] A. Jackson, J. Kidd, Performance, Learning and Heritage, Centre for Applied Theatre Research, Manchester, 2008.

[11] J. Kidd, Gaming for affect: Museum online games and the embrace of empathy, 2015. URL: https://intellectdiscover.com/content/journals/10.1386/jcs.4.3.414_1. doi:https://doi.org/10.1386/jcs.4.3.414_1.

[12] P. Bruijn, Bridges to the Past: Historical Distance and Multiperspectivity in English and Dutch Heritage Educational Resources, Ph.D. thesis, Unpublished, 2014. doi:10.13140/RG.2.2.27514.29123.

[13] A. Witcomb, Understanding the role of affect in producing a critical pedagogy for history museums, Museum Management and Curatorship 28 (2013) 255–271. URL: https://doi.org/10.1080/09647775.2013.807998. doi:10.1080/09647775.2013.807998.

[14] J. Kidd, With new eyes i see: embodiment, empathy and silence in digital heritage interpretation, International Journal of Heritage Studies 25 (2019) 54–66. URL: https://doi.org/10.1080/13527258.2017.1341946. doi:10.1080/13527258.2017.1341946.

[15] Y. Du, Confirmation bias in generative ai chatbots: Mechanisms, risks, mitigation strategies, and future research directions, 2025. URL: https://arxiv.org/abs/2504.09343. arXiv:2504.09343.

[16] E. Lopez-Lopez, C. M. Abels, D. Holford, S. M. Herzog, S. Lewandowsky, Generative artificial intelligence–mediated confirmation bias in health information seeking, Annals of the New York Academy of Sciences 1550 (2025) 23–36. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC12412720/. doi:10.1111/nyas.15413.

[17] V. Berthet, P. Teovanović, V. de Gardelle, A common factor underlying individual differences in confirmation bias, Scientific Reports 14 (2024) 27795. URL: https://www.nature.com/articles/s41598-024-78053-7. doi:10.1038/s41598-024-78053-7.

[18] K. A. Costabile, S. Madon, Downstream effects of dispositional inferences on confirmation biases, Personality and Social Psychology Bulletin 45 (2019) 557–570. URL: https://doi.org/10.1177/0146167218789624. doi:10.1177/0146167218789624.

[19] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch,

N. Schiefer, D. Yan, M. Zhang, E. Perez, Towards understanding sycophancy in language models, 2025. URL: https://arxiv.org/abs/2310.13548. arXiv:2310.13548.

[20] C. M. Abels, E. Lopez-Lopez, J. W. Burton, D. L. Holford, L. Brinkmann, S. M. Herzog, S. Lewandowsky, The governance i& behavioral challenges of generative artificial intelligence's hypercustomization capabilities, Behavioral Science I& Policy 11 (2025) 22–32. URL: https://doi.org/10.1177/23794607251347020. doi:10.1177/23794607251347020.

[21] N. Simon, Fighting for inclusion, 2015. URL: http://museumtwo.blogspot.com/2015/09/fighting-for-inclusion.html.