Discovery Machines

written by Jeremy Huggett | 29/08/2023

Adapted from the original by Brian J. Matis (CC BY-NC-SA 2.0)

Michael Brian Schiffer is perhaps best-known (amongst archaeologists of a certain age in the UK at least), for his development of behavioural archaeology, which looked at the changing relationships between people and things as a response to the processual archaeology of Binford et al. (Schiffer 1976; 2010), and for his work on the formation processes of the archaeological record (Schiffer 1987). But Schiffer also has an extensive track record of work on archaeological (and behavioural) approaches to modern technologies and technological change (e.g., Schiffer 1992; 2011) which receives little attention in the digital archaeology arena, in part because despite his interest in a host of other electrical devices involved in knowledge creation (e.g., Schiffer 2013, 81ff) he has little to say about computers beyond observing their use in modelling and simulation or as an example of an aggregate technology constructed from multiple technologies and having a generalised functionality (Schiffer 2011, 167-171).

In his book *The Archaeology of Science*, Schiffer introduces the idea of the 'discovery machine'. In applying such an apparatus,

> … the investigator makes a series of substitutions for one of the components or undertakes a series of new interactions. In this manner, the ever-changing apparatus, *still possessing the same basic structure*, yields one new effect after another, some of which may also become discoveries. A discovery machine, then, generates a *discovery cascade*. (Schiffer 2013, 192, emphasis in original).

Such characteristics sound very applicable to a computer, though Schiffer does not make the connection himself. His use of the concept of a 'discovery machine' is broadly equivalent to its use

elsewhere; for instance, Huff (2011) describes the development of the telescope and its subsequent use by Galileo to study the moon, the satellites of Jupiter, etc. as a discovery machine, a development also discussed by Schiffer (2013, 192) along with Hooke's development of the microscope. But Schiffer also broadens out the idea of 'discovery machines' to include technologies in earlier societies such as cooking pots and distillation apparatus (2013, 194). In many respects, discovery machines are akin to cognitive artefacts – human-made objects employed as a means of assisting us to perform a cognitive task, able to represent, store, retrieve, and manipulate information – that I discussed in a digital archaeology context a few years ago (Huggett 2017). In both cases, the technological tool complements human investigation through extending and supporting the labour of research and discovery; it does not replace the human component in the process (e.g., see Schmidt & Loidolt 2023).

But what if a discovery machine could be constructed that replaced the human creative and cognitive effort? It has long been thought that this would be impossible, although some have looked towards the creation of an Artificial General Intelligence or High Level Machine Intelligence ultimately resulting in the arrival of the Singularity where technology comes to exceed human intelligence (e.g., Vinge 1993; 2008 – and see, for example, Brin et al. 2013). Today we might be forgiven for thinking that such a singularity was at least closer given the latest crop of large language model (LLM) Generative Pretrained Transformers (GPT) such as ChatGTP, LLaMa, and Bard which have captured the headlines of late, along with similar tools beginning to become embedded in our search tools and office software. Their appearance of intelligence is misleading: as Levine (2023, 51) has observed, for example, they are not designed to give factually correct answers, but to arrange a set of words that is consistent with human syntax by sequentially selecting the most probable word to follow in a string. Hence critics have famously labelled them 'stochastic parrots', "haphazardly stitching together sequences of linguistic forms it has observed in its training data, according to probabilistic information about how they combine, but without any reference to meaning" (Bender et al. 2021, 617). Not all agree with this characterisation of LLMs as a glorified if complex cut-and-paste machine (e.g., Arkoudas 2023) but nevertheless accept that such systems have no actual understanding of the content they create which is why much of the criticism surrounding them, once past the hype, has begun to highlight the way that they easily invent 'facts', propagate misinformation, have difficulty drawing inferences and hence establishing causation (as opposed to being very good at identifying correlation), are incapable of distinguishing between truth and falsehood in either their data or their responses, and can therefore 'hallucinate' improbable or impossible scenarios (e.g., Arkoudas 2023, 4ff; Denning 2023; Levine 2023).

One key attraction of these systems is their ability to analyse large volumes of disparate data, but they are sensitive to bias and imbalances within their training data and data quantity is no substitute for quality. For instance, recent work by Alex Reisner (2023) has shown that a dataset called 'Books3' was used in the training of LLaMA and other generative-AI programs. Aside from the copyright implications, from an archaeological perspective the 'Books3' dataset included much of the opus of Erich von Däniken, and we might also hazard that works by the likes of Graham Hancock may likely appear. While (for all we know) they may be outnumbered by mainstream archaeology texts in the dataset, what does this imply for the archaeological 'knowledge' of these tools? How quickly does pseudoarchaeology surface in archaeological enquiries? Spenneman (2023) recently investigated the extent to which archaeological literature may have been used in training ChatGTP and concluded that, in the absence of detailed knowledge of the training data, even the appearance of valid references in its responses could not be taken as indicating the actual text behind the

reference had been used as it was unable to quote from the texts themselves, and many of the references were likely derived from Wikipedia. Agapiou & Lysandrou (2023, 4081-3) found that while high-level responses to inquiries concerning remote sensing in archaeology were generally accurate, in-depth, more comprehensive responses required a number of follow-up questions; they also comment on the lack of information about sources. Elsewhere, ChatGTP has been caught manufacturing bogus references and summarizing their non-existent content, underlining it as "an application for *sounding like an expert*, not for *being an expert*" (Davis 2023)

One reply might be to move away from these generalised, generic systems and train them with specifically archaeological data, narrowing down the training set to data which relate to archaeological questions rather than incorporating a host of irrelevant material which creates non-archaeological 'noise' in the system. This is already happening elsewhere, with specialised GPT's appearing for medical diagnosis and computer programming, for example. Attractive as this might sound – and some archaeologists are already working in this area – nevertheless it overlooks the fundamental shortcomings of the way these models operate, as outlined above. But even their underlying 'knowledge' remains problematic – they capture a particular world view with the biases in the data propagated through the system, and we have yet to establish clear principles for how to create or maintain appropriately balanced training data, or indeed how those biases and other constraints might affect the system. Keeping humans in the loop (HITL) may be one way of ensuring the integrity of the system – refining the training data and checking the outputs, altering the parameters so as to improve the quality of responses, but in most cases, HITL is seen as part of tuning and improving the system, with the ultimate objective being to replace the human component with a fully automated accurate and reliable process. Further, simply adding the human at the end as well as at the outset doesn't address the problem of the intervening black box – how is an archaeologist to evaluate an outcome without knowledge of how it was arrived at? The human conception of an archaeological problem will not match the system's representation of it, and the system's mode of reasoning will not be easily understandable by the human. And how can any knowledge produced by the system be trusted if that same system, inscribed with unseen biases and hidden perspectives, is capable of invention, confabulation, or hallucination while seeking to justify its conclusions?

Schiffer (2013, 192) writes of discovery machines generating 'discovery cascades', but it is equally true that these digital discovery machines generate 'risk cascades' (or 'data cascades', e.g., see Sambasivan 2022) at every level, from the collection of data, labelling the data, selecting the data, cleaning the data, selecting the computational model, selecting the training data, training the model, evaluating the model, and finally applying the model. Errors or imbalances at any stage are compounded at each subsequent stage, cascading through the system in ways that are often unexpected by the user and certainly unrecognised by the system itself. We therefore have an ethical responsibility to understand the underlying assumptions at each stage in order to properly evaluate the system outcomes. However, the thrill of conversing with the machine, using a generalised model more often developed by others, the way the system seeks to justify or evidence its conclusions, and the obfuscation of the underlying data and reasoning mechanisms, makes this challenging to achieve but all the more important if we are to ensure the human component remains core to the creation of archaeological knowledge.

References

Agapiou, A., & Lysandrou, V. 2023. Interacting with the Artificial Intelligence (AI) Language Model ChatGPT: A Synopsis of Earth Observation and Remote Sensing in Archaeology. *Heritage*, 6(5), 4072–4085. doi: 10.3390/heritage6050214

Arkoudas, K. 2023. ChatGPT is no Stochastic Parrot. But it also Claims that 1 is Greater than 1. *Philosophy & Technology*, 36(3), 54. doi: 10.1007/s13347-023-00619-6

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp.610–623. New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3442188.3445922

Brin, D., Broderick, D., Bostrom, N., Chislenko, A., Hanson, R., More, M., Neilsen, M. & Sandberg, A. 2013. A Critical Discussion of Vinge's Singularity Concept. In M. More & N. Vita-More (Eds.), *The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future*, pp. 395–417. Chichester, UK ; Malden, MA: John Wiley & Sons, Ltd. doi: 10.1002/9781118555927.ch37

Davis, B. 2023. We Asked ChatGPT About Art Theory. It Led Us Down a Rabbit Hole So Perplexing We Had to Ask Hal Foster for a Reality Check. *ArtNet News* (March 2, 2023) https://news.artnet.com/art-world/chatgpt-art-theory-hal-foster-2263711

Denning, P. J. 2023. The Smallness of Large Language Models. *Communications of the ACM*, 66(9), 24–27. doi: 10.1145/3608966

Huff, T.E. 2011. *Intellectual Curiosity and the Scientific Revolution: A Global Perspective.* New York: Cambridge University Press.

Huggett, J. 2017. The Apparatus of Digital Archaeology. *Internet Archaeology*, 44. doi: 10.11141/ia.44.7

Levine, E. V. 2023. Cargo Cult AI. *Communications of the ACM*, 66(9), 46–51. doi: 10.1145/3606946

Reisner, A. 2023. Revealed: The Authors Whose Pirated Books Are Powering Generative AI. *The Atlantic* https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/

Sambasivan, N. 2022. All equation, no human: The myopia of AI models. *Interactions*, 29(2), 78–80. doi: 10.1145/3516515

Schiffer, M.B. 1976. *Behavioral archeology*. New York: Academic.

Schiffer, M.B. 1987. *Formation processes of the archaeological record*. Albuquerque: University of New Mexico Press.

Schiffer, M. B. 1992. *Technological perspectives on behavioral change*. Tucson: University of Arizona Press.

Schiffer, M. B. 2010. *Behavioral Archaeology: principles and practice*. London: Equinox

Schiffer, M.B. 2011. *Studying technological change: A behavioral approach*. Salt Lake City: University of Utah Press.

Schiffer, M.B. 2013. *The Archaeology of Science: Studying the Creation of Useful Knowledge*. Heidelberg: Springer International Publishing. doi: 10.1007/978-3-319-00077-0

Schmidt, P., & Loidolt, S. 2023. Interacting with Machines: Can an Artificially Intelligent Agent Be a Partner? *Philosophy & Technology*, 36(3), 55. doi: 10.1007/s13347-023-00656-1

Spennemann, D. H. 2023. What has ChatGPT read? The origins of archaeological citations used by a generative artificial intelligence application. *arXiv*. doi: 10.48550/arXiv.2308.03301

Vinge, V. 1993. The Coming Technological Singularity: How to Survive in the Post-Human Era. *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, pp.11–22. Cleveland, OH: National Aeronautics and Space Administration. Retrieved from https://archive.org/details/NASA_NTRS_Archive_19940022855/mode/2up

Vinge, V. 2008. Signs of the singularity. *IEEE Spectrum*, 45(6), 76–82. doi: 10.1109/MSPEC.2008.4531467