Draft version – accepted for PSA 2024 contributed paper proceedings

Can Generative AI Produce Novel Evidence?

Donal Khosrowi (Leibniz University Hannover)
donal.khosrowi@philos.uni-hannover.de

Finola Finn (University of Luxembourg, Leibniz University Hannover)
finola.finn@uni.lu

Abstract

Researchers in archaeology explore the use of generative AI (GenAI) systems for reconstructing destroyed artifacts. This paper poses a novel question: can such GenAI systems generate *evidence* that provides new knowledge about the world or can they only produce *hypotheses* that we might seek evidence *for*? Exploring responses to this question, the paper argues that 1) GenAI outputs can at least be understood as *higher-order evidence* (Parker 2022) and 2) may also produce de novo *synthetic evidence.*

Keywords: Artificial intelligence, Generative AI; Evidence; Historical Science; Archaeology; Computer Simulation; Epistemology

1 Introduction

Artificial intelligence (AI) systems, including generative AI (GenAI), play ever larger roles across the sciences: they are used to make novel discoveries, e.g., of proteins, drugs, or materials (Jumper et al. 2021; Sourati and Evans 2023); to identify new concepts and equations in physics (Iten et al. 2020; Udrescu et al. 2020; Wu and Tegmark 2019); and to suggest new hypotheses, ideas, research questions or experiments (Krenn et al. 2023; Melnikov et al. 2018). These increasingly extensive roles played by AI put foundational concepts we use to understand and structure scientific pursuits under pressure (Löhr 2023; Hopster and Löhr 2023). For instance, what does it mean to be a scientific 'discoverer' (Clark and Khosrowi 2022)? Are AI systems like AlphaFold only 'tools' that humans use, or can they exhibit attributes such as 'autonomy' or scientific 'understanding' (Gonzalez Barman et al. 2023) which we consider essential to the role of 'discoverer' or 'researcher'?

While use cases of AI in physics, chemistry and biology attract increasing attention by philosophers, there are also underexplored emerging uses of AI in the historical sciences (i.e. history, archaeology, anthropology), where researchers explore the use of GenAI for reconstructing partially destroyed manuscripts and artifacts (Navarro et al. 2022; 2023) and other tasks. Turning attention to these uses, this paper draws out a novel conceptual disruption regarding how we should understand the *outputs* of GenAI systems: can GenAI systems *generate* evidence that provides genuinely new knowledge in the way that, say, finding new material evidence can? Or can they only produce hypotheses, which may give us reasons for pursuit, but ultimately are the kind of thing that we require evidence *for*? Call this the *evidence question.* Like other conceptual disruptions caused by AI, the evidence question does not have a straightforward answer and highlights substantial uncertainty around how we should apply the concept of 'evidence'. The issues this raises are not merely terminological ones but have epistemic and methodological import for practicing researchers. Classifying an output as 'evidence' rather than a 'hypothesis' confers information about it: existing norms may trigger different expectations, attitudes and actions as appropriate in relation to an output.

Beyond putting the evidence question on the map, this paper also explores potential responses to it. We first consider related debates in the philosophy of computer simulation, where scholars such as Wendy Parker (2022) have elucidated whether simulation systems, such as those used in climate science, can provide (new) evidence for claims about the earth's climate system. Drawing on this debate, we argue that GenAI systems can at least provide *higher-order evidence* in Parker's sense, i.e. evidence that other evidence for a claim about the world exists. We also explore a more ambitious argument, according to which GenAI systems can produce de novo *synthetic evidence,* which could be epistemically on par with traditional forms of evidence, such as material evidence or expert judgment. We suggest they do so by performing pattern recognition-type inferences to yield outputs that provide genuinely new knowledge to agents who lack the ability to make those same inferences.

The discussion is organized as follows. Section 2 outlines the emerging use of GenAI in the historical sciences. Section 3 sharpens the evidence question. Section 4 explores debates in the philosophy of computer simulation and sketches the sequential arguments that GenAI can at least produce higher-order evidence, as well as, possibly, synthetic evidence. Section 5 concretizes why we may take GenAI outputs seriously. Section 6 concludes.

2 Generative AI in the Historical Sciences

A central challenge for researchers in the historical sciences is that the 'record' of historical evidence, e.g. manuscripts or artifacts such as pottery, is eroded – both figuratively and literally. Not everything survives, and what does is often incomplete or broken. A standard activity in getting a handle on the past, e.g. to stratify archaeological sites, entertain larger inferences about chronology, locating or classifying ethnic groups, or inferring trade patterns, is hence to reconstruct what was from what remains. Reconstructing partially destroyed artifacts, e.g. to better determine relevant morphological or textural features, is currently often performed by hand, which is resource intensive, can further deteriorate remaining fragments, and cannot deal with fragments that are missing (Navarro et al. 2023). Dealing with such and similar challenges, there is a rich tradition in the historical sciences, especially in archaeology, to recruit techniques and technologies from adjacent fields (Wylie 2000), e.g. for sensing and scanning, or, in computational archaeology, involving machine-learning based technologies. For instance, Navarro et al. (2023) develop a GenAI system based on generative adversarial network (GAN; Goodfellow et al. 2014) called *IberianVoxel*, which reconstructs 3D-models of broken Iberian pottery artifacts. GANs consist of a coupled generator and discriminator architecture – in Navarro et al.'s case, the generator produces 3D-voxel geometries and the discriminator 'judges' whether the geometries produced by the generator look like they were drawn from the data distribution of scanned real artifacts on which it is trained. After a period of adversarial training, the GAN is evaluated, including by surveying domain experts to assess reconstruction quality. The authors report that "[…] archaeologists judge that IberianVoxel generated a correct Iberian style from an initial fragment, and also consider that the reconstructed pottery is between Good and Very Good." (2023, 5839) and conclude their system is "[…] very helpful for exploring and designing automatic procedures to aid experts with the pottery completion task" (ibid., 5833).

Systems such as IberianVoxel are first steps on a likely trajectory towards more advanced systems permitting finer-grained inferences, especially as GenAI technologies become cheaper to train. Extrapolating along this trajectory, let us imagine a stylized toy case inspired by IberianVoxel to draw out the central question of this paper more clearly. Consider *AlphaPot*, an imagined GenAI system that has been trained on a very large dataset consisting of images and corresponding high-quality 3D-scans of a wide range of pottery artifacts in various states of decay. AlphaPot is trained to reconstruct masked/corrupted features of an input, e.g. parts of a 3D-mesh from a real scanned artifact that are intentionally corrupted (e.g. either by generating synthetic data based on real artifacts that simulate fragmentation or by physically breaking artifacts and then re-scanning them) and the system is forced to predict how the uncorrupted artifact would have looked like. Assume we are impressed with AlphaPot's performance on held-out test data: it accurately reconstructs broken artifacts for which the ground truth geometry is known. Imagine now that we venture to provide a reconstruction R of a novel, partially destroyed artifact A, for which the ground truth is unknown. A is missing pieces that haven't been recovered, but are believed to be essential to classifying A's likely origin or function. R, let us assume, is a plausible looking, high-resolution 3D-mesh containing fine-grained morphological features that would significantly aid a domain expert (or, for that matter, another AI system) in telling when and where A originated.

3 What's Going on Here, Epistemically?

The key disruption motivating this paper is now clearly in view: what's going on here, epistemically? Has AlphaPot generated a *hypothesis* or made a *prediction*? Or has it generated *evidence*, providing experts with genuinely new knowledge about how A looked like when it was still intact? Understood as a mere *hypothesis*, R is the kind of thing that might give us reasons for pursuit, seeking further evidence to support that R is indeed what A looked like when it was still intact. By contrast, understood as *evidence*, R might already, by itself, entitle us to form novel, justified beliefs about A, or to make knowledge claims.

In making progress on the evidence question, we need to agree on a benchmark first. A standard Bayesian conception of evidence requires only that a token of evidence E has the capacity to increase the posterior probability we assign to a hypothesis H (say, a claim about how A looked like when it was intact) relative to some background theory T (Bovens and

Hartmann 2003). It is easy to imagine that R has *some* such capacity, but that is not a very interesting insight. Other, functional conceptions of evidence articulate what *role* evidence plays. Here, we are sympathetic to accounts that consider evidence as always being 1) *of* something, 2) *for* something, and 3) *to* someone, relative to a theory of evidence (Hacking 2006; Martini 2021). For instance, a freshly excavated artifact A is evidence *of* something, e.g. the fact that pottery of A's kind was made, used or traded at site S; evidence *for* something, e.g. an inferred claim that pottery of A's kind was produced in P but ended up at S through a trade route; and evidence *to* someone who has a theory of evidence T and relevant background knowledge K to tell what A can be evidence *of* and *for*. Beyond following such structured conceptions, the subsequent discussion will remain largely uncommitted to specific *philosophical* accounts of evidence. Instead, we find it more productive to consider evidential practices in the historical science and think about what existing benchmark types of evidence we could compare GenAI outputs to. What could such benchmarks be? It seems unhelpful, for instance, to think that observing A when it was still intact, or freshly made is a relevant benchmark for clarifying whether R is evidence for claims about A or wider inferences that draw on beliefs about A. Rather, the historical sciences have a distinct preference for primary, material sources, e.g. artifacts and manuscripts that are close (causally, spatially, temporally, by provenance) to the phenomena of interest. Relevant benchmarks to address the evidence question could hence be, for instance, a highly similar, intact artifact B found in the same stratum at the same site; pertinent text, illustrations or tools bearing on the likely morphological features of A; or expert judgment that joins up available background theory and knowledge together with primary material evidence of the aforementioned kinds in an inference. The evidence question is sharper now: could AlphaPot's outputs be considered evidence comparable to these benchmarks, e.g. other, material evidence like B that could licence an analogical inference that 'A would have probably looked like B when it was intact'; or expert judgment that joins various resources together to yield, say, a rendition or description of what A would look like, were it still intact?

4 Yes, but What Kind of Evidence Is It?

The answer that we want to explore here is: yes, GenAI systems like AlphaPot have the capacity to generate *synthetic evidence* that provides genuinely new knowledge about the world. What could an argument for such a thesis look like? A first pass could build on familiar successes of using AI systems for inferential tasks in science, like AlphaFold 2.0 (Jumper et al. 2021). Specifically, at training, these systems 1) latch onto information, especially high-dimensional and distributed correlational information or patterns, in training data, and 2) learn a model, i.e. an abstract representational space encoding relevant features and a corresponding function $f(\cdot)$ within that space, which maps inputs to outputs in a way that minimizes empirical risk (at least locally). At inference, given an input (e.g. a scan of a partially destroyed artifact), 3) such models generate outputs that yield accurate reconstructions of the input, as governed by $f(\cdot)$. This kind of story would touch on guarantees in machine learning (e.g. Cybenko 1989) and statistical learning theory (e.g. Vapnik 2000; Bargagli Stoffi et al. 2022) to explain notable successes of machine learning systems, e.g. in latching onto complex, subtle, and distributed patterns that escape human attention, such as in skin cancer classification or protein structure prediction (Jumper et al. 2021), or in successfully learning novel high-dimensional representations (e.g. word or image embeddings) that can be used for text and image synthesis, as demonstrated by GenAI systems like ChatGPT or StableDiffusion (Rombach et al. 2022).

This story, while somewhat compelling, is still too simple, however. Here, we focus on concerns arising from related debates in the philosophy of scientific models and computer simulation. In this space, philosophers have tried to understand whether models and simulations can provide genuinely new knowledge about the world and, if so, how (e.g. Parker 2022; Beisbart 2012). In a nutshell, sceptics about the epistemological significance of models and simulations point out that these tools only help us recognize the consequences of knowledge that we already possess, e.g. assumptions (e.g. equations) and initial conditions (e.g. measurements, parameterizations). These consequences can at most be evidence in the sense that they provide new information to agents who are not able to, or simply did not, derive those same consequences given the same assumptions and initial conditions. But they would not be evidence to a more ideal agent who would already recognize these consequences under some form of inferential closure. So, while observation and experimentation allow us to gather *new* experience (Beisbart 2012, 245), models and simulations don't add anything new to the table; though they do help limited epistemic agents better see what's already on the table.

What does this mean for the evidence question? Parker summarizes the consequences of the sceptical view on computer simulation as follows: "If computer simulation is at bottom an attempt to calculate the implications of a set of modelling assumptions, then simulation results […] seem to be *predictions rather than evidence*; they are the kind of thing we might seek evidence for […]." (Parker 2022, 1522; emphasis added) On such a view, the outputs of GenAI systems like AlphaPot are predictions, or, more generally, hypotheses. They might alert us to possibilities for how an artifact may have looked like, and may give us reasons for pursuing these hypotheses by means of bringing evidence to bear on them; but they are not to be taken as evidence that already, by itself, licenses knowledge claims or underwrites downstream inferences, such as about trade taking place between different communities.

Filling the space between more extreme views that either consider simulation results evidence, or deny that they can ever be, Parker offers a finer-grained view to characterize what simulations provide to agents. Specifically, Parker argues that simulation outputs can be *higher-order evidence*: they can be evidence E that other evidence E' for a hypothesis H exists. Specifically, such higher-order evidence can help agents obtain genuinely new knowledge of the world if they 1) either don't have access to E', or else, 2) lack the background knowledge needed to understand how E' bears on H. So, while simulations "[…] do not provide information about the world that goes beyond that which is already implicit in their assumptions, particular epistemic agents—including even scientists and engineers using simulation models—might still gain genuinely new knowledge of the world via simulation." (Parker 2022, 1522)

Parker's view offers a useful backstop for thinking about GenAI outputs. At the very least, they seem able to figure as higher-order evidence. A 3D reconstruction of a broken artifact A from a suitably validated system like AlphaPot provides new knowledge about specific artifacts to agents who either don't have access to the training data[1] E' that bear on the reconstructive query about A, or else lack the background knowledge to understand how E' bears on questions about A. This is a useful insight already, but it also seems interesting to explore whether GenAI systems could ever provide more than 'just' higher-order evidence.

4.1 More Than Higher-Order?

What might GenAI systems be doing, or capable of, that goes beyond what simulation systems do? A central difference seems to be that GenAI systems can exhibit higher degrees of independence, which allows them to perform computations that instantiate *inferences* of a kind that simulation systems don't instantiate. Specifically, simulation systems in the climate sciences are built based on highly concrete antecedent understanding of the physics equations describing aspects of the earth's climate system (background knowledge), are parameterized according to our best understanding of key parameters and known/understood aspects of the phenomena involved, and are calibrated using data regarding the earth's climate system. Together, these inputs substantially constrain the behaviors of simulating systems.

GenAI systems exhibit comparatively higher independence because they are not as tightly constrained. There are no accepted equations that describe, say, the 'grammar' of Iberian pottery. Nor, for lack of such equations, are there measurements that a GenAI system is parameterized with. In short, there is no developed body of background knowledge that goes into building GenAI systems (at least in unsupervised/self-supervised regimes), nor would our existing background knowledge permit building systems in a way that mirrors the strategies behind building climate simulation systems. Rather, the very purpose of machine learning approaches to problems such as those faced in reconstructing artifacts is often to *extract* pertinent background knowledge from data, e.g. to find a function $f(\cdot)$ that usefully captures a joint distribution and can be used to perform successful inferences. For this enterprise to be successful, GenAI systems must exhibit considerable degrees of freedom to 'settle' on representational spaces, representations, and input-output relationships that are 1) predictively useful, 2) not up to human agents to identify, and 3) possibly inaccessible to humans by other means (e.g. visual inspection) and 4) novel to humans. GenAI systems hence harbor the capacity for a special kind of novelty in their outputs. Unlike simulation systems, they can generate *synthetic evidence*, i.e. evidence E that is not only psychologically novel to agents who lack other evidence E' or background knowledge K, but is novel to agents who do not possess the same *inferential abilities* to extract pertinent knowledge K (e.g. of $f(\cdot)$) from the same training data. Such abilities are different from *computational* abilities to derive *implications* of equations and initial conditions. They are more akin to the ability to

---

[1] In particular, they might lack access to the *information* contained in that data, e.g. regularities about the 'grammar' of Iberian pottery.

'recognize' that such-and-such is a good way to represent or compress data, or that such-and-such is a successful (i.e. error-minimizing) way to 'fill in the blanks' of a reconstructive query.

On the narrative presented here, GenAI systems bring inferential abilities to the table that simulation systems don't. But why should this lead us to conclude that they can *generate* evidence that provides genuinely new knowledge to agents? Couldn't, or shouldn't, we still maintain that the relevant information with bearing on reconstructive queries 'resides in' the training data that GenAI systems are trained on?[2] This would bring us back to understanding GenAI systems as, at most, providing higher-order evidence in Parker's sense and conclude that no evidence that is novel over and above whatever is contained in these data is generated.

A good way to explore how GenAI systems can provide novelty beyond higher-order evidence is to think about patterns. A standard success narrative of machine-learning based inference alluded to above is centrally tied to the ability to identify patterns, including subtle and distributed ones, and exploit them for inference. But what is a pattern, anyway? This is just another conceptual disruption that ML systems press us to confront with greater care. Here, it is useful to distinguish two general types of views we can take on patterns: *ontic* views and *epistemic* views. On the first, a pattern is constituted by a collection of material facts about the world that may be distributed across entities. A pattern, on this view, is always 'there', even without a mind to recognize and exploit it for inference. We are not aware of work defending such a view. On an epistemic view (cf. Dennett 1991; McAllister 2010), patterns come into existence only through an epistemic *agent* that recovers it, including, say, by devising an ontology of entities and features ranging over a domain (e.g. pots, fractures, materials, textures), making efforts to describe and represent these entities and features within that domain in an abstract way (e.g. material or shape types), and exploring how these representations hang together, e.g. causally or probabilistically, at that abstract representational level. On such a view, a pattern is instantiated by, refers to, and supervenes on, concrete material things, but ultimately resides at an abstract representational level. If we find such an epistemic view compelling, then this allows that GenAI systems, like other epistemic agents, can perform inferential activities that *bring patterns into existence*.[3] It is this ability, which sets GenAI systems apart from simulating systems in that they produce an output, based on the ability to infer patterns from data, that is novel to an agent who does not possess similar inferential abilities.

5 Strictures on Synthetic Evidence

We now have a sketch of an argument for the claim that GenAI systems like AlphaPot may produce *synthetic evidence*, i.e. evidence E that provides genuinely new knowledge about the

---

[2] This concern also flags a version of the *problem of old evidence* for Bayesian confirmation theory; see e.g. Sprenger (2015) for a discussion.

[3] Of course, we must be careful here in navigating a conceptual landscape rich in anthropomorphic pitfalls. Terms like 'recognizing', 'using', and so on, must not be taken to suggest that GenAI systems literally have mental states or cognitive abilities often associated with these terms.

world to agents who do not possess the same inferential abilities to recover E from primary evidence E’ as the system that produced E. But when can we expect GenAI systems to produce *good* synthetic evidence? As researchers are exploring use-cases of LLMs in history (Hutson et al. 2024), for instance to ‘ventriloquize’ the voices of the past through LLMs trained and/or fine-tuned on historical text corpora to enable researchers to ‘query’ past societies or individuals, there is a real risk of low-cost bogus AI-driven science. Here, we outline some potential virtues that GenAI systems may exhibit, if designed and deployed responsibly. These virtues help better understand the conditions under which we may reasonably hope these systems to make valuable epistemic contributions.

1) **Comprehensiveness:** GenAI systems are good at processing and ‘drawing’ on large amounts of rich data, which is relevant when patterns are distributed across large numbers of entities and different data modalities.
2) **Sensitivity:** ML systems are known to usefully latch onto subtle, distributed patterns, especially in quantitative data, that are often not accessible to human perception.
3) **Probabilism**: ML-based inference is probabilistic. Outputs are sampled from a whole modeled joint distribution. This often means that other possibilities for an output are not discarded by a system, but remain, or can be made, available to investigators.
4) **Mechanicity:** outputs are often (near-) repeatable from the same inputs, and so GenAI systems can be subjected to systematic intervention, allowing investigators to understand how outputs depend on inputs. For instance, they may upsample rare input types (e.g., by using synthetic training data to induce more variation regarding specific artefact types) and gauge whether outputs change for specific query types.
5) **Theory-freedom/-agnosticism**: especially in unsupervised or self-supervised learning regimes, GenAI systems organize data somewhat independently of existing theory, working against unhelpful forms of theory-ladenness of observation.
6) **Complexity:** universal function approximation theorems (e.g. Cybenko 1989) and statistical learning theory (Vapnik 2000; Bargagli Stoffi et al. 2022) provide (probabilistic) guarantees for specific system-types to successfully approximate arbitrarily complex input-output relationships under suitable conditions. This is important as there are no good reasons to believe that, say, the ‘grammar’ of, say, Iberian pottery, i.e. the ‘rules’ that govern the joint distribution of morphological features of Iberian pottery artifacts, is easily captured by simple, human-expressible functions.
7) **Granularity:** GenAI systems perform inference at multiple levels, including at fine-grained pixel- or voxel levels that may not be salient to human investigators. Such systems are hence not as susceptible as humans to latch exclusively onto patterns or analogies obtaining at higher, more salient levels of analysis, e.g. inferring that artifact A probably had inscription S because, B, C, D, who look morphologically similar, do.

Of course, specific GenAI systems are not guaranteed to exhibit any of these virtues to significant degrees; only well-engineered systems may. Moreover, many of the candidate virtues outlined here can turn into vices if the properties they track are expressed too strongly – think of theory-freedom or comprehensiveness that could lead a system to consider irrelevant or misleading information when there are good theoretical reasons not to. Spelling out a contextualist virtue epistemology for GenAI systems in science is arguably a larger project that will require more space, which is why the virtues sketched here should only provide some early inspiration rather than a sketch of a full-fledged account of what GenAI systems may bring to the table. That said, it seems promising to explore such an account in articulating answers to the evidence question.

6 Conclusions

This paper puts an important new question about the role of generative AI (GenAI) systems in the sciences on the map. The *evidence question* asks: can GenAI systems *generate* evidence that provides agents, including experts, with genuinely new knowledge about the world? Focusing on the historical sciences, where researchers explore the use of GenAI systems to reconstruct partially destroyed manuscripts and artifacts to learn about the past, we argued that it is currently unclear whether we should understand the outputs produced by these systems as mere *hypotheses* or as *evidence*, where the former may give researchers reasons for pursuit and seeking out further evidence, and the latter may already licence knowledge claims about the world and figure directly in underwriting further inferences. Given this conceptual and practical uncertainty, we sketched how we may understand GenAI outputs as *synthetic evidence*, i.e. evidence that can provide agents, including experts, with genuinely new knowledge about the world. In a nutshell, compared to even our most sophisticated simulating systems, such as those used in climate science, GenAI systems do not merely compute the consequences of a conjunction of background knowledge and existing evidence, but acquire and deploy inferential abilities to produce outputs that are evidence to agents who lack those same inferential abilities, which may include even our best domain experts. The scope of this argument sketch is narrow: it applies, for now, only to the emerging uses of GenAI in the historical sciences discussed here. But zooming out, the evidence question may also extend to a range of other domains that explore the utility of GenAI for advanced inferential tasks (e.g. drug and materials discovery). For philosophers of science this is good news, inviting us to help characterize and resolve the methodological disruptions affecting emerging scientific practices and contribute to development of sound methodologies involving GenAI.

References

Bargagli Stoffi, Falco J., Gustavo Cevolani, and Giorgio Gnecco. 2022. "Simple Models in Complex Worlds: Occam's Razor and Statistical Learning Theory." *Minds & Machines* 32:13–42 (2022). https://doi.org/10.1007/s11023-022-09592-z

Beisbart, Claus 2012. "How can computer simulations produce new knowledge?" *European Journal for Philosophy of Science 2*(3):395–434.

Bovens, Luc, and Stephan Hartmann. 2003. *Bayesian Epistemology*. Oxford: Oxford University Press.

Clark, Elinor, and Donal Khosrowi. 2022. “Decentring the discoverer: how AI helps us rethink scientific discovery.” *Synthese* 200(463). https://doi.org/10.1007/s11229-022-03902-9

Cybenko, George. 1989. “Approximation by Superpositions of a Sigmoidal Function.” *Mathematics of Control, Signals, and Systems* 2:303-314.

Dennett, Daniel C. 1991. “Real Patterns.” *Journal of Philosophy*, LXXXVIII, 27-51.

Gonzalez Barman, Kristian, Sascha Caron, Tom Claassen, and Henk de Regt. 2023. “Towards a Benchmark for Scientific Understanding in Humans and Machines”. Preprint at https://doi.org/10.48550/arXiv.2304.10327

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. “Generative Adversarial Nets.” *Advances in Neural Information Processing Systems* 27.

Hacking, Ian. 2006 The emergence of probability: a philosophical study of early ideas about probability, induction and statistical inference. Cambridge University Press, Cambridge

Hutson, James, Paul Huffman, Jeremiah Ratican, 2024. "Digital resurrection of historical figures: A case study on Mary Sibley through customized ChatGPT." *Metaverse* 4(2). doi: 10.54517/m.v4i2.2424

Hopster, Jeroen, Philip Brey, Michael Klenk, Guido Löhr, Samuela Marchiori, Björn Lundgren, and Kevin Scharp. 2023. Conceptual Disruption and the Ethics of Technology. In Ibo van de Poel, Lily Frank, Julia Hermann, Jeroen Hopster, Dominic Lenzi, Sven Nyholm, Behnam Taebi, and Elena Ziliotti (Eds.), *Ethics of Socially Disruptive Technologies: An Introduction* (pp. 141-162). Cambridge, UK: Open Book Publishers. https://doi.org/10.11647/OBP.0366.06

Hopster, Jeroen and Guido Löhr. 2023. “Conceptual Engineering and Philosophy of Technology: Amelioration or Adaptation?” *Philosophy & Technology* 36: 70. https://doi.org/10.1007/s13347-023-00670-3

Iten, Raban, Tony Metger, Henrik Wilming, Lidia del Rio, and Renato Renner. 2020. “Discovering physical concepts with neural networks.” *Physical Review Letters* 124: 010508. https://doi.org/10.1103/PhysRevLett.124.010508

Jiang, Harry H., Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. 2023. “AI Art and its Impact on Artists.” In P*roceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23),* Association for Computing Machinery, New York, NY, USA, 363–374. DOI: https://doi.org/10.1145/3600211.3604681

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596:583-589. https://doi.org/10.1038/s41586-021-03819-2

Krenn, Mario, Lorenzo Buffoni, Bruno Coutinho, Sagi Eppel, Jacob Gates Foster, Andrew Gritsevskiy, Harlin Lee, Yichao Lu, João P. Moutinho, Nima Sanjabi, Rishi Sonthalia, Ngoc Mai Tran, Francisco Valente, Yangxinyu Xie, Rose Yu, and Michael Kopp. 2023. "Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network." *Nature Machine Intelligence* 5:1326-1335. https://doi.org/10.1038/s42256-023-00735-0

Löhr, Guido. 2023. "Conceptual disruption and 21st century technologies: A framework." *Technology in Society* 74:102327. https://doi.org/10.1016/j.techsoc.2023.102327

Martini, Carlo. 2021. "What "Evidence" in Evidence-Based Medicine?" *Topoi* 40: 299–305. https://doi.org/10.1007/s11245-020-09703-4

McAllister, James W. 2010. "The Ontology of Patterns in Empirical Data." *Philosophy of Science* 77(5):804–814. doi:10.1086/656555

Melnikov, Alexey. A., Hendrik P. Nautrup, Mario Krenn, and Hans H. Briegel. 2018. "Active learning machine learns to create new quantum experiments." *PNAS* 115(6):1221-1226. https://doi.org/10.1073/pnas.1714936115

Navarro, Pablo, Celia Cintas, Manuel Lucena, José Manuel Fuertes, Rafael Segura, Claudio Delrieux, and Rolando González-José. 2022. "Reconstruction of Iberian ceramic potteries using generative adversarial networks." *Scientific Reports* 12:10644. https://doi.org/10.1038/s41598-022-14910-7

Navarro, Pablo, Celia Cintas, Manuel Lucena, José Manuel Fuertes, Antonio Rueda, Rafael Segura, Carlos Ogayar-Anguita, Rolando González-José, and Claudio Delrieux. 2023. "IberianVoxel: automatic completion of iberian ceramics for cultural heritage studies." In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI '23). Article 647, 5833–5841. https://doi.org/10.24963/ijcai.2023/647

Sprenger, Jan 2015. "A new solution to the problem of old evidence." *Philosophy of Science, 82*(3):383–401.

Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Björn Ommer. 2022. "High-Resolution Image Synthesis with Latent Diffusion Models." arXiv. DOI: https://doi.org/10.48550/arXiv.2112.10752.

Parker, Wendy S. 2022. "Evidence and Knowledge from Computer Simulation." *Erkenntnis* 87: 1521–1538. https://doi.org/10.1007/s10670-020-00260-1

Sourati, Jamshid and James A. Evans. 2023. "Accelerating science with human aware artificial intelligence." *Nature Human Behaviour* 7:1682-1696. https://doi.org/10.1038/s41562-023-01648-z

Udrescu, Silviu-Marian, Andrew Tan, Jiahai Feng, Orisvaldo Neto, Tailin Wu, and Max Tegmark. 2020. “AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity.” *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*: 4860–4871.

Wu, Tailin, and Max Tegmark. 2019. “Toward an artificial intelligence physicist for unsupervised learning”. *Physical Review E*, 100(3): 033311. https://doi.org/10.1103/PhysRevE.100.033311

Vapnik, Vladimir N. 2000. *The nature of statistical learning theory*. New York: Springer.

Wylie, Alison. 2000. "Questions of Evidence, Legitimacy, and the (Dis)Unity of Science." *American Antiquity* 65(2):227-237. https://doi.org/10.2307/2694057