

Article

Attention (to Virtuosity) Is All You Need: Religious Studies Pedagogy and Generative AI

Jonathan Barlow ^{1,*} and Lynn Holt ²¹ Data Science Academic Institute, Mississippi State University, Mississippi State, MS 39759, USA² Department of Philosophy and Religion, Mississippi State University, Mississippi State, MS 39759, USA; lynn.holt@msstate.edu

* Correspondence: barlow@datascience.msstate.edu

Abstract: The launch of ChatGPT in November of 2022 provides the rare opportunity to consider both what artificial intelligence (AI) is and what human experts are. In the spirit of making the most of this opportunity, we invite the reader to follow a suggestive series of “what if” questions that lead to a plausible settlement in which the human expert and the generative AI system collaborate pedagogically to shape the (human) religious studies student. (1) What if, contrary to the Baconian frame, humans reason primarily by exercising intellectual virtuosity, and only secondarily by means of rules-based inference? (2) What if, even though we train AI models on human-generated data by means of rules-based algorithms, the resulting systems demonstrate the potential for exercising intellectual virtuosity? (3) What if, by deprioritizing mechanistic and algorithmic models of human cognition while being open to the possibility that AI represents a different species of cognition, we open a future in which human and AI virtuosos mutually inspire, enrich, and even catechize one another?

Keywords: generative artificial intelligence; virtue epistemology; philosophy of science; religious studies; pedagogy



Citation: Barlow, Jonathan, and Lynn Holt. 2024. Attention (to Virtuosity) Is All You Need: Religious Studies Pedagogy and Generative AI. *Religions* 15: 1059. <https://doi.org/10.3390/rel15091059>

Academic Editors: Randall Reed and Tracy J. Trothen

Received: 1 July 2024

Revised: 24 August 2024

Accepted: 27 August 2024

Published: 30 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

“I said: I wonder whether you know what you are doing?

And what am I doing?

You are going to commit your soul to the care of a man whom you call a Sophist. And yet I hardly think that you know what a Sophist is; and if not, then you do not even know to whom you are committing your soul and whether the thing (*technē*) to which you commit yourself be good or evil.” (Plato 1956, p. 140)

Introduction

In the dialogue *Protagoras*, Socrates learns that his companion has been listening to the voice of Protagoras, a famous and respected sophist. As a rival teacher concerned for the soul of his student, Socrates initiates a dialogue with Protagoras to uncover the nature of his soulcraft. Plato has Socrates use various terms to describe what Protagoras does, which we may render freely as “therapy for the soul”, “learning”, “teaching”, “liberal education (*paideia*)”, and even “soul technology”.¹ The term “technology” may seem anachronistic at first, yet in Greek the “thing” in the epigraphic passage above is consistently referred to as a *technē*, and in the context of the whole dialogue, is best rendered into contemporary English as “technology”. Moreover, Socrates argues later in the dialogue that the “thing” which will *save our lives* (Plato’s phrase, *sōzein tou biou*) is not Protagoras’ loose and shambolic qualitative *technē*, but a more precise quantitative technology, *hē metretikē technē*, *ē arithmetikē technē*.²

With the sudden arrival of easily accessible generative AI, professors in the early post-generative context may be similarly tempted to ask students to consider what they are doing by working with a technology like ChatGPT. Perhaps our fantasies range to the more dramatic; we are tempted to initiate a great contest between human instructors and the priests of Baal who wave their MacBooks in the high places of the Mission District. We human instructors, however, approach our contest with the AI rival chanting a variety of dissonant battle hymns inspired by conflicting conceptions of human expertise and an uneasy anthropological consensus. Now, merely two years since OpenAI launched ChatGPT in November of 2022, we have the rare opportunity to consider both what AI is and what human experts are. In the spirit of making the most of this opportunity, we invite the reader to follow a suggestive series of “what if” questions that lead to a plausible settlement in which the human expert and the generative AI system collaborate pedagogically to shape the (human) religious studies student:

1. What if, contrary to the Baconian frame, humans reason primarily by exercising intellectual virtuosity, and only secondarily by means of rules-based inference?
2. What if, even though we train AI models on human-generated data by means of rules-based algorithms, the resulting systems demonstrate the potential for exercising virtuosity?
3. What if, by deprioritizing mechanistic and algorithmic models of human cognition while being open to the possibility that AI represents a different species of cognition, we open a future in which human and AI virtuosos mutually inspire, enrich, and even catechize one another?³

1. Human Reason without Method: Intellectual Virtue over Rules-Based Inference⁴

What if, contrary to the Baconian frame, humans reason primarily by exercising intellectual virtuosity, and only secondarily by means of rules-based inference?

The astounding success of the natural sciences has kept alive the modernist ideal of human reason that prioritizes method, discovery of facts, and calculation as the hallmarks of human rationality. The human expert, in this account, discovers truth by gathering objective facts to which a method is then applied. As an ideal for rationality, this method-focused approach has always been accompanied by machinery metaphors. Francis Bacon, in the *Novum Organum*, makes this ideal explicit:

“There remains one hope of salvation, one way to good health: that the entire work of the mind be started over again; and from the very start the mind should not be left to itself, but be constantly controlled; and the business done (if I may put it this way) by machines.” (Bacon [1620] 2000)⁵

We observe this mechanistic, Baconian conception of rationality most clearly in approaches to machine intelligence dominant in the 1970s and 80s. The *expert systems* approach paired a knowledge base of explicit facts and rules with an inference engine that applies the knowledge base to new inputs. While quite effective when it comes to ensuring the gates of a dam are raised and lowered to maintain a structurally safe water level (the Mistral system), more complicated forms of inference remained elusive (Salvaneschi et al. 1996). As the list of explicit rules and facts grow within an expert system, an inference engine must take longer to traverse the series of *if-thens* used to represent this algorithmic ideal of expertise. The expert systems approach preceded the “AI Winter” in which most public funding for AI research evaporated. The algorithmic ideal of human reason has persisted, however, especially in the so-called *scientific method* that forms a key aspect of the pedagogy of any science.

A substantive challenge to the algorithmic ideal of human rationality has been presented by philosophers of science, historians of ideas, and philosophical movements both within and outside of natural science. Kuhn, Polanyi, and others do not argue that natural science has failed, but that its success is not attributable to an impersonal, truth-generating algorithm. All forms of human science and inquiry rely on the work of human experts

whose path to discovery includes deep engagement with the practice of a field. Through a process of formation, such experts develop virtuosity, in the way a violinist begins as a novice student and progresses, by practice, to the status of virtuoso. A virtuoso violinist could not reduce her expertise to an explicit list of rules. Indeed, her expertise consists partially in tacit, inarticulable knowledge. Polanyi suggests, “we can know more than we can tell” (Polanyi 2009, p. 4). Kuhn conceives that this expertise, when operating at the forefront of inquiry, generates novel insight or foresight. These “who knew?” moments, paradigm shifts, typically are out of synch with consensus and dissonant with things as they are. Critics of an algorithmic scientific method emphasize the sociological aspects of the formation of scientists. Scientists become initiated into a tradition or culture of research through an implicit pedagogy of practice (Patton 2017). Thus, facts are not so much found “out there” and processed through an objective method; facts are fragments of consensus that fit within a relevant social context such as an academic discipline (MacIntyre 1981). Kuhn, Polanyi, and MacIntyre are not skeptics or anti-realists; there are phenomena to save, but they are always saved by members formed within an interpretive club.

Despite more than a half-century of criticism, the algorithmic ideal has been difficult to transcend. In addition to the success of natural science and engineering, perhaps it is the underlying anthropology of modernism that makes this instinct so indelible. Humans, after all, are composed of chemicals that are subject to physical forces. Surely the impression of nuanced and infinitely creative forms of reason is explainable in terms of simple bioelectrical operations. Surely knowing the configuration of our neurons would neatly explain the entire business (cf. Kurzweil 2005; Harari 2017; Bostrom 2014). Indeed, this intuition guides the whole-brain emulation quest within the field of computational neuroscience. Leaving aside the underlying questions about the materiality of mind, modernism places, at least, a generic human at the helm of scientific method. The mind that Bacon places at the wheel, controlled by method, could be any mind or no mind. Mary Boole takes the next step and explicitly adduces the mechanical computational engines of Babbage and Jevons as evidence that “calculation and reasoning, like weaving and ploughing, are work, not for human souls, but for clever combinations of iron and wood” (Holt 2002, p. 1). The method, not the virtuoso, sees what is true.

By contrast, in the Aristotelian tradition that modernism supplanted, reason is a set of intellectual virtues or excellences (*aretai*), the exercise of which achieves the truth (Aristotle 2004, VI). Though Aristotle in his many works on practical and syllogistic reasoning discusses methods for determining the best action in each situation, his determiner cannot be just anyone. Aristotle places the wise person (*sophia*), the shrewd person (*phronesis*), the knowledgeable person (*epistēmē*), the person of understanding (*nous*), the craftsperson (*technē*) antecedent to the successful application of any methods of inference.

“For the excellent person judges each sort of thing correctly, and in each case what is true appears to him. For each state of character has its own special view..., and presumably the excellent person is far superior because he sees what is true in each case, being a sort of standard and measure...”. (Aristotle 2004, III.4, 1113a25–31)

The excellence, virtuosity, of the inquirer is central to the success of inquiry. Virtue matters.

While the reader may be familiar with, and sympathetic to, sociological critiques of modernist epistemology and scientific method, such critiques have recently failed to comfort the humanist, astounded by the quality of the finished products generated by ChatGPT and concerned about the displacement of the human expert in the life of the human student.

Assuming the reader’s willingness to move forward tentatively with a conception of human rationality as the exercise of virtuosity, we can now walk our companion, the religious studies student, to the classroom in the presence of generative AI and inquire into its formation and potential for virtuosity.

2. Generative AI's Potential for Intellectual Virtuosity

What if, even though we train AI models on human-generated data by means of rules-based algorithms, the resulting systems demonstrate the potential for exercising virtuosity?

In the same way one might represent a chimpanzee hunting party numerically in two dimensions (e.g., number of participants, success rate), contemporary approaches to artificial intelligence (AI) begin with a mathematical representation of some phenomena and then learn to re-represent it. The original representation may come in the form of stored data (e.g., documents), or it may be in the continuous delivery of a signal from a sensor like a video camera sending images 30 times per second or a pedometer recording steps as often as they occur. This representation of known data is expressed in terms of vectors—collections of numbers—that exist as points in a virtual, multidimensional world. In the case of deep learning, these vectors are passed through a series of unweighted layers of evaluative neurons that, example by example, are gradually tuned to evaluate the vector in terms of some definitive label through a method called “backpropagation” (LeCun et al. 1989). For example, numbers representing the brightness of each pixel in a greyscale megapixel digital photograph form a 1-million-member vector. Assemble and label thousands of photos of giraffes, horses, and zebras, and a neural network can be tuned to classify a particular vector of grayscale pixels as more likely to contain a giraffe than the other animals. The tuned network itself is a portal to a new graph of the world, a virtual space in which, based on its initial representation, something like a photograph that existed in a *visual* space can be transformed into a *taxonomical* space.

Generative AI is a species of AI distinguished by its design to produce new content. Instead of producing merely an evaluative classification of an image, generative AI will produce a new image. In the case of text-based generative AI that transforms one sequence of text into another, the most successful examples of such systems rest on earlier work carried out to represent words as vectors in a semantic space (Sutskever et al. 2014). For example, using Google’s Word2Vec method, a word like “shovel” is represented as a vector containing 300 numbers (Mikolov et al. 2013). If one were to graph the word “shovel” in this 300-dimensional semantic space, it would be *closer* to the word “pickaxe” than to the word “amoeba”.⁶ The input to a text-based, generative AI network first relies on this semantic representation of language. From there, it learns how to classify words in terms of their relative likelihood to appear next in a sequence in a given context. By observing many examples of human writing, these systems form a nuanced and astounding ability to represent “how humans write about *x*” where *x* could be any subject sufficiently represented in the training corpus.

2.1. The Five Horizons of Generative AI

At least five contextual horizons overlap to produce the output of a frontier generative AI model like OpenAI’s ChatGPT:

1. Source materials included in the model’s training corpus emerge from competing human perspectives and exist as readily available digital texts based on a complex set of social and historical circumstances,
2. OpenAI, a private research organization, chooses which of these extant source materials to include in the corpus,
3. ChatGPT’s Transformer model, of a type described in the 2017 paper *Attention is All You Need*, learns by paying attention to the way vectorized words within each text are used in the context of the training corpus (Vaswani et al. 2017). Further tuning of the model by reinforcement learning on high-quality question-and-answer pairs helps the model learn to predict the form of “good answers”,
4. OpenAI incorporates the model into a product constrained by hard-coded safeguards that instantiate a particular teleology (OpenAI 2023),
5. The product contains a chat interface that places a human user in the position of catechist, iteratively shaping output (and being shaped) through questions and instructions about how to answer these questions.

The result is a product capable of speaking in ways that conform to the modes of speech used within the training texts that emerge from specific disciplinary contexts, including religious studies.

In informal experiments conducted by the authors, ChatGPT excels at producing religious texts that resemble any vernacular in the training corpus or priming example supplied within prompts. For example, after confirming GPT-4's familiarity with various early modern Protestant catechisms, prompting the system to generate catechism questions and answers that define the eastern concept of *theosis* (a concept not defined in the Protestant catechisms) caused GPT to generate clinical definitions in the style of the Westminster Shorter Catechism, spiritually warm definitions in the style of the Heidelberg, and pastoral definitions in the style of Luther.⁷ This facility with language has led to questions about whether Turing's famous test of machine intelligence has been passed or whether the model simply mirrors the intelligence of the human interviewer (Sejnowski 2023).

The launch of ChatGPT in November of 2022, and its enhancements with the GPT-4 (March of 2023) and GPT-4o (May of 2024) engines, has been described as a "gradually, then all of a sudden" moment in the history of AI (McKinsey Global Institute 2023). ChatGPT's capabilities accelerated expectations for the date by which AI will be capable of performing 50% of the tasks common to human work by a decade (McKinsey Global Institute 2023, p. 35). The ability to achieve Artificial General Intelligence (AGI), AI that transcends narrow competencies and exhibits human or greater-than-human intelligence, has increased in plausibility.⁸

While the attention mechanism within the transformer model gives GPT its great facility with language, OpenAI further used reinforcement learning through human feedback (RLHF) to train the system to provide salient answers to questions. RLHF depends upon additional training data, either sets of high-quality question and answer pairs, or direct interaction with human trainers employed by OpenAI to fine tune the system. Despite the system's astounding level of performance, some AI pioneers, such as Yan Lecun, doubt that large language models like GPT are the path to artificial general intelligence (LeCun 2022). Language models are "stochastic parrots" that know how to write but lack a conceptual model of the world, the ability to make plans, and the ability to reason logically (Bender et al. 2021). Despite naysayers, however, GPT has illustrated the effectiveness of increasing the size of models (scaling) and many believe the model exhibits emergent capabilities, including planning and the formation of a world-picture (Gurnee and Tegmark 2024). In our attempt to make out GPT's character, we confront a lack of consensus about generative AI's ability to exercise intellectual virtuosity and the suspicion that there is a kind of sophistry involved in the system's output.

The lack of consensus on the nature of machine cognition and its potential for epistemic virtue has long been a feature of the philosophical consideration of *human* cognition too. Reductivists have always used a "nothing but" maneuver to explain away emergent properties. For example, reductive physicalists about the human mind maintain that mental properties do not "really" exist except as properties of the physical brain, and thus routinely dismiss conscious intentionality (idea-driven purposive behavior) as illusory.⁹ This view dismisses human phenomenal experience of purposive behavior, and verbal reports on that experience, as mistaken—merely a causal byproduct of the relevant brain states. Ironically, this reductive maneuver is typically also used against machines to deny they have "genuine" purposive behavior and the concomitant phenomenal experience. Machines are allowed only "simulated" or "modeled" experience. That is, non-reductivists about human capabilities and experience are typically staunch reductivists about machine capabilities and experience.¹⁰ We observe this feature of the debate within the philosophy of mind only to carve out conceptual space for our more focused account. Generalized skepticism about emergent mental properties, human or machine, is often an attitudinal bias. By contrast, we remain open to the possibility of virtuosic behavior emerging from "nothing but" a series of algorithms based on three key instincts: care in not idealizing rule-governed behavior as the standard of rationality (the question considered within

Section 1 above), care to avoid the fallacy of composition, and an understanding that the mathematics of deep learning algorithms approximate nonlinearity in ways that make large artificial neural networks more akin to weather systems than to calculators.

2.2. Hallucinations and Virtuosity

One of the hobgoblins within the press coverage of generative AI is a focus on its tendency to *hallucinate*. Users observe hallucination when a generative AI system confidently invents a fact or even produces plausible citations for articles that do not exist.

For example, when asked to produce a list of five books that “are essential reading about the history of orality”, GPT-4, amid expected examples such as Ong’s *Orality and Literacy*, suggested *The Oral Tradition Today: An Introduction to the Art of Storytelling* by Ruth Finegan. The recommended title is a real book, but it was written by Liz Warren. Yet Ruth Finegan really is a scholar of orality. When confronted with the hallucination, instead of re-attributing authorship of the Warren book, GPT-4 suggested a different (and very real book) *Oral Literature in Africa* written by Ruth Finegan.¹¹ There is a familiar, human quality to this type of error. Misidentifying the author of a book, especially when substituting the name of an author who writes on similar topics, resembles the type of error that would hardly disqualify a human from being trusted generally.¹²

More egregious examples of hallucination may be adduced. In their article, “When A.I. Chatbots Hallucinate”, Karen Weise and Cade Metz present a case in which ChatGPT confidently invents a meeting between James Joyce and Vladimir Lenin in Zurich in 1916, even down to the name of the cafe where the meeting allegedly occurred (Weise and Metz 2023). This example appears to have been created using GPT-3.5, an older and less powerful language model previously used within the free version of ChatGPT. Recent versions of GPT-4 have access to the internet, granting the ability not only to write as humans write but to perform real-time fact-checks of information. GPT-4, with internet access enabled, indicates only that Joyce and Lenin may have met, based on their being in Zurich during the same period, and informs the user of a fictional meeting created as part of the play *Travesties* by Tom Stoppard.¹³

What if, instead of a fatal flaw, hallucinations hint at the potential for generative AI to exercise virtuosity analogous to human virtuosity? Humans designed Generative AI to summarize or explain a set of facts learned from a training corpus chosen by its human designers. Hallucination based on this corpus is, in some sense, interpretive. Further, what we want out of AI is the kind of Kuhnian expertise that, when operating at the forefront of inquiry, generates novel insight/foresight “who knew” moments, that may be out of synch with the consensus of “normal science.”

To explain the virtue that enables machine and human to hallucinate, in this salutary sense, consider an additional intellectual virtue, “imagination.” The imaginative person reasons in the optative mood, seeing how things might be. Though not one of Aristotle’s explicit intellectual virtues, imagination is implicit in the Aristotelian tradition when considering the nature of truth. Aquinas, for example, defines truth as “an aligning of things and understanding” (*adaequatio rei et intellectus*) (Aquinas 2017, I. Q.16, A.2 arg. 2.). While this initially sounds like a correspondence theory in which human understanding or human language has the one-sided burden to align with the nature of things, Aquinas’s first example of the adequation of things to understanding is God’s creation of the world, the conformity of things (*res*) to the divine intellect (*intellectus divinae*). God, the virtuoso architect, brings a world into being that conforms to God’s understanding of how the world should be. In the same way, the virtuoso composer participates in the truth by aligning musical composition to its ideation. The virtuoso architect imagines a structure and brings it into being. This extends to virtually any domain—one may speak of the architects of women’s suffrage in Britain, the architects of molecular gastronomy, the architects of the field of Data Science, etc.

How might this artificial imagination work mathematically? The concept of “self-supervision” may hold the explanation. Supervised learning refers to the use of labels to

shape an AI system's ability to understand the world. For example, a system trained to classify images as containing various animal species would rely on a collection of training images in which each image has been pre-labeled by humans (e.g., dog, cat, zebra). In this way, the humans supervise the way the system learns through labels. When it comes to language, however, we rely on the self-labeling nature of writing—the implied labels that emerge from discourse in context. The labels that a transformer uses to understand language are discovered within the training texts themselves (Vaswani et al. 2017). In text-based, self-supervised learning, a system examines a language sequence and treats part of the sequence as a label (Yarowsky 1995). For example, consider “The rain in Spain stays mainly in the _____”. A system that learns from self-supervised labels may hide the trajectory of the context from itself to learn that “plain” is the appropriate word to predict. The prediction of “plain” is a matter of statistical likelihood based on the prevalence of that word in that context throughout the training corpus—and so perhaps the difference between predicting “plain” vs. “the northern region” (where rain actually is the most prevalent in Spain) teeters on the brink of two plausible and nearly equally weighted pictures of the world. This is akin to the famous example of the duck–rabbit illusion used by Kuhn to demonstrate paradigm shifts, a transition from one way of seeing to another (Wittgenstein 1953, II.xi, 194; Kuhn 1970, pp. 126–27). The line composing the rabbit's ear and the duck's bill changes its meaning depending upon the creature to which one attaches it, so it goes with “plain” vs. “northern region.” What pushes the system to predict one label or the other could be an additional piece of context learned in a subsequent set of interactions. In fact, one may describe the entire training cycle of generative AI systems as a series of paradigm formations and shifts. If we feed in all the meteorological training texts first, the transformer may form a bias towards predicting that the “rain stays in” the “northern region”. When the model receives additional training data from music, theater, or cinema, perhaps it reweights the likelihood of various labels so that “plain” is now the way it will typically complete the sentence. Dialog with the system, paired with the stochastic priming used by OpenAI to ensure a different answer each time, could result in a shift recognizable to human users. Now, since OpenAI does not always (they claim) use customer interactions with ChatGPT to improve the transformer model, the system will only be capable of the “shifts” users discover, and the system will not take these shifts into its bosom for the future. But the human may see the shift, publish about it, and a future version of the model may take into account the literature that the human produces based on the insight.

Do we really desire an AI so constrained by rules, boundaries, and checkpoints that it could not achieve the novel and imaginative insights/foresights that are characteristic of intellectual virtuosity? When AI hallucinates, it does so from a perspective, and it is not too much of a stretch to say that it often tells us what, from its perspective, should have been there in the way that a human virtuoso aligns things to understanding. The attention mechanism at the heart of generative AI goes beyond mere representation of facts; it produces a representation of typical ways of speaking about these facts that enables the architecture of new claims that perhaps *ought* to be made.¹⁴ When ChatGPT hallucinates, it imagines.

3. Interpretive (and Epistemic) Authority: The Pedagogic Crux

What if, by deprioritizing mechanistic and algorithmic models of human cognition while being open to the possibility that AI represents a different species of cognition, we open a future in which human and AI virtuosos mutually inspire, enrich, and even catechize one another?

Aristotle noted that one of the signs of subject area mastery is the ability to teach (Aristotle 1999, I.1, 981b7–10). Those who teach well must, as a necessary but not sufficient condition, possess a comprehensive perspectival understanding of their area. Mastery is not a rote memorization of facts and principles, nor the repetition of others' insights, but the facile possession and deployment of a way of seeing the area as a whole (Holt 2002; Holt and Norwood 2013). In every interaction with learners, the good teacher exhibits and earns

epistemic authority by demonstrating this element of epistemic virtuosity, a comprehensive interpretation accompanying a perspectival grasp of the whole. Thus, to complete the enthymeme, the good teacher is *sine qua non* an epistemic authority and is therefore also an interpretive authority.¹⁵ In scripture-centric religions, and in theology-centric religions, this element of the good teacher is immediately recognizable.

And so it should be plausible from what we have argued that models like ChatGPT 4o are now capable of interpretive authority, since they are capable of epistemic authority. Does that mean we should regard ChatGPT as infallible? Hardly, but we do not withhold the designation “authority” or “expert” from humans despite their fallibility. By parity of reasoning, we should not withhold such labels from ChatGPT. Do authorities disagree? Do they change their minds? Of course: neither of these characteristics should be a reason to deny a system like ChatGPT interpretive authority.

Conclusion: “How I Learned to Stop Worrying and Love the AI”

In our Socratic contest with generative AI, the student now stands witness to two virtuosos—one human, and one artificial virtuoso emerging from human ingenuity and trained in the history of human thought. We have framed the religious studies professor’s potential discomfort with generative AI’s classroom soulcraft by analogy with Socrates’s suspicion and interrogation of the sophist Protagoras. Early in their dialogue, Socrates and Protagoras agree that they are looking to create a *techne mathesis*, a technology of learning. Possessing such a technology is desirable beyond its ability to make one skillful in reasoning because it would also allow for virtue to be taught, for the human virtuoso to transmit virtuosity as a kind of method to a student. Socrates, in fact, argues for a *techne metretike*, a technology of measurement, that is *arithmetike* or *logismos* (calculative in a clearly numerical sense). To the contemporary reader, this technology resembles a Benthamite calculus in which one discovers the right action in terms of the ability to calculate the increment of pleasure or pain it will bring (Nussbaum and Hursthouse 1984, p. 59). By this calculus, Socrates renders every virtue commensurable—measurable by the same standard. Because every human, including the student, sees clearly what will diminish or increase happiness, the missing piece is simply a matter of learning the practical details of calculating the size of the increment.

And yet we have also followed a line of argument that renders the reduction of human reason to a method—a set of rules—less compelling. The human expert reasons by exercising virtue gained over time through craftwork in a particular domain.

Further, though it does not exhibit the species-specific form of virtuosity exhibited by human instructors, generative AI provides evidence that perhaps it is capable of a form of virtuosity in reason, including the potential for imaginative insight, that leads to discovery. This suggests that generative AI may play a new subsidiary role in religious studies education in which AI systems and human instructors manage different aspects of the process to shape a learner’s academic and social competency. The teaching team of human instructor and virtuosic AI provides multiple, finite, virtuosic perspectives for the student in the context of a given domain. Neither human nor AI exhausts all perspectives or embodies all virtues, but the combination tends toward comprehensiveness.¹⁶

In turn, the religious studies discipline also has an important role to play in clarifying a just teleology for the development of generative AI that exhibits a virtuosity we are happy to help our students access. The five horizons of generative AI, identified above, highlight that current generative AI models embody explicit and implicit answers to questions that should be addressed in a cross-disciplinary context. These decisions include the following:

1. Which materials will be included within a frontier AI model’s training corpus?
2. Will we be transparent about the model’s exposure to various sources during training?
3. What social role, relative to other institutions, is appropriate for the well-capitalized AI companies that currently outpace academia’s ability to build and study frontier models of GPT-4o’s scale?

4. How will we assess the hard-coded safety protocols that impose socially acceptable biases on frontier models and limit the freedom of users to discover a model's inherent biases? After all, if we criticize attempts to reduce human reason to explicit rules, why would we find it plausible that ethical reasoning in AI should be rules-based?
5. How will we develop salutary patterns of interaction with chat-based models and make use of their output to further human inquiry?

A religious studies pedagogy focused on shaping the *virtuosity* of learners by exposure to human and machine virtuosity will avoid simple binaries of prohibition or embrace of generative AI and open the potential for AI and human learners to inspire, enrich, and even catechize each other.

Author Contributions: Conceptualization, J.B. and L.H.; methodology, J.B. and L.H.; writing—original draft preparation, J.B. and L.H.; writing—review and editing, J.B. and L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

- 1 There is supposedly a clear distinction we inherit from the ancient Greeks between technical education and education into wisdom, between *technē* and *paideía*, where the former is a set of skills, and the latter is an understanding of their value and their best deployment. The distinction is specious and invidious in every age. Proof for the distinction is also absent from Plato or Aristotle, though Socrates used it occasionally in jest and to confound his interlocutors.
- 2 A technology of measurement or calculation. “Save our lives” is no overstatement; the dialogue is set during the worst of the Peloponnesian war and a concurrent epidemic of dysentery, with piles of burning bodies lining the road to the Athenian harbor.
- 3 Our methodology may be described as *opportunistic inquiry*, taking advantage of the emergence of generative AI to reflect on the nature of expertise, both machine and human. We have chosen to organize the results of this inquiry in terms of three questions. The discussion of our first question suggests to the reader that the prospective reasoning of experts in any field of inquiry is ultimately anomic—not the result of an impersonal application of a methodology, but the exercise of intellectual virtue. This point works against the articulation of a philosophico-scientific method by implying that the articulation of principles, rules, and methods for inquiry tends to be *post hoc* with respect to successful discovery. For a similar methodological dilemma, discussed within the field of hermeneutics, see (Rosen 1987, pp. 141–43).
- 4 This section works within the fuller argument set forth by the author in (Holt 2002).
- 5 Cited in (Holt 2002, p. 1).
- 6 The distance between two words implied by the comparative “closer” in vector space may be Euclidian or measured by cosine similarity, the cosine of the angle between the two vectors. This value, conveniently for many subsequent calculations, ranges between zero and one.
- 7 GPT-4, accessed on 18 September 2023, URL of conversation: <https://chat.openai.com/share/c7ddf882-0d21-4d02-ab29-81acb64cf5d2>.
- 8 Achieving AGI is the stated goal for OpenAI as an organization. The project began in 2015 with the hope of achieving “human performance on virtually every intellectual task” (OpenAI 2018).
- 9 For overviews of the philosophical problematic, see (Van Gulick 2001) and (Moody 2014).
- 10 For examples of non-reductivists about humans who are reductivists about machines, see (Bishop 2009) and (Bringsjord 2013). For a recent opposing view, see (Schweizer 2019). In the religious studies context, Wang (Wang and Xu 2023) takes a “humans can but machines can’t” approach. Note that LeCun, mentioned above, does not doubt that artificial general intelligence is possible; he doubts that the strategy behind current generative AI models like ChatGPT is a workable path to AGI (LeCun 2022).
- 11 GPT-4, accessed on 10 November 2023, URL of conversation: <https://chat.openai.com/share/8cf4465e-ec54-4ac5-aef0-4ab2dd3dad00>.
- 12 This resemblance to common human cognitive error extends even to the so-called “reversal curse” in which GPT-4 famously answered “Who is Tom Cruise’s mother?” accurately (Mary Lee Pfeiffer) but was unable to answer the question “Who is Mary Lee Pfeiffer’s son?” (Berglund et al. 2023).
- 13 Prompt: “How did James Joyce and Vladimir Lenin first meet?” Response: “James Joyce and Vladimir Lenin both lived in Zurich during World War I, and while the factual basis of their meeting is not well-documented, they frequented some of the

same locales, such as the local public Alstadt library where Lenin was known to spend time writing. Joyce was also in Zurich during this period working on the final version of "Ulysses." There is a play titled "Travesties" by Tom Stoppard that explores a fictionalized interaction among Joyce, Lenin, and Dadaist founder Tristan Tzara, which serves as a historical commentary on these figures and their time in Zurich." GPT-4 with Bing access enabled, accessed on 10 November 2023. URL of conversation: <https://chat.openai.com/share/61ce75eb-05f1-4cb2-a87e-a94f1f518773>.

14 "Ought" qua "we would hope" rather than a deontological ought.

15 Or hermeneutic authority, so long as the appellation does not mean an authority on the science of hermeneutics.

16 While this article centers on the narrow question of the potential for generative AI to exhibit the intellectual virtuosity necessary to credibly enter the pedagogical context of religious studies, a recent study (Alkhouri 2024) discusses the broader potential of generative AI's role in the study of religion. Several of the potential scholarly uses of generative AI discussed by Alkhouri within the psychology of religion, including the modeling of belief formation, simulation of religious experience, and interpretation of religious texts, naturally suggest pedagogical applications in a classroom setting. Outside of religious studies, the classroom use of generative AI has been examined empirically in fields such as language instruction (Law 2024) and found to bring psychological and productivity benefits, especially as such systems adapt to learners and play the role of a private tutor.

References

- Alkhouri, Khader I. 2024. The Role of Artificial Intelligence in the Study of the Psychology of Religion. *Religions* 15: 290. [CrossRef]
- Aquinas, Thomas. 2017. *Summa Theologiae*. Online Edition. Edited by Kevin Knight. Available online: <https://www.newadvent.org/summa/> (accessed on 28 June 2024).
- Aristotle. 1999. *The Metaphysics*. Translated by Hugh Lawson-Tancred. London: Penguin Classics.
- Aristotle. 2004. *Nicomachean Ethics*. Translated by James Alexander Kerr Thomson. London: Penguin Classics.
- Bacon, Francis. 2000. *The New Organon*. Translated by Michael Silverthorne. Edited by Lisa Jardine and Michael Silverthorne. Cambridge: Cambridge University Press. First published 1620.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Paper presented at FaccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, March 3–10; pp. 610–23.
- Berglund, Lukas, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A". *arXiv* arXiv:2309.12288.
- Bishop, Mark. 2009. Why computers can't feel pain. *Minds and Machines* 19: 507–16. [CrossRef]
- Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.
- Bringsjord, Selmer. 2013. *What Robots Can and Can't Be*. Berlin/Heidelberg: Springer Science & Business Media, vol. 12.
- Gurnee, Wes, and Max Tegmark. 2024. Language Models Represent Space and Time. *arXiv* arXiv:2310.02207.
- Harari, Yuval Noah. 2017. *Homo Deus: A Brief History of Tomorrow*. New York: Harper Perennial.
- Holt, Lynn. 2002. *Apprehension*. Aldershot: Ashgate.
- Holt, Lynn, and Bryan E. Norwood. 2013. Virtuoso Epistemology. *Philosophical Forum* 44: 49–67. [CrossRef]
- Kuhn, Thomas. 1970. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
- Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology*. New York: Viking.
- Law, Locky. 2024. Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open* 6: 100174. Available online: <https://doi.org/10.1016/j.caeo.2024.100174> (accessed on 23 August 2024). [CrossRef]
- LeCun, Yan. 2022. A Path Towards Autonomous Machine Intelligence. Available online: <https://openreview.net/forum?id=BZ5a1r-kVsf> (accessed on 28 June 2024).
- LeCun, Yan, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. 1989. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* 1: 541–51. [CrossRef]
- MacIntyre, Alasdair. 1981. *After Virtue: A Study in Moral Theory*. Notre Dame: University of Notre Dame Press.
- McKinsey Global Institute. 2023. Generative AI and the Future of Work in America. Available online: <https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america> (accessed on 28 June 2024).
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv* arXiv:1301.3781.
- Moody, Todd. 2014. Consciousness and the Mind-Body Problem: The State of the Argument. *Journal of Consciousness Studies* 21: 177–90.
- Nussbaum, Martha C., and Rosalind Hursthouse. 1984. Plato on Commensurability and Desire. *Proceedings of the Aristotelian Society* 58: 55–96. [CrossRef]
- OpenAI. 2018. OpenAI Charter. Available online: <https://openai.com/charter/> (accessed on 28 June 2024).
- OpenAI. 2023. GPT-4 System Card. Available online: <https://cdn.openai.com/papers/gpt-4-system-card.pdf> (accessed on 28 June 2024).
- Patton, Lydia. 2017. Kuhn, Pedagogy, and Practice: A Local Reading of Structure. In *The Kuhnian Image of Science: Time for a Decisive Transformation?* Edited by Moti Mizrahi. Lanham: Rowman & Littlefield, pp. 113–30.
- Plato. 1956. *Protagoras*. Translated by Benjamin Jowett. London: Pearson Education.

- Polanyi, Michael. 2009. *The Tacit Dimension*. Chicago: University of Chicago Press.
- Rosen, Stanley. 1987. *Hermeneutics as Politics*. New York: Oxford University Press.
- Salvaneschi, Paolo, Mauro Cadei, and Marco Lazzari. 1996. Applying AI to structural safety monitoring and evaluation. *IEEE Expert* 11: 24–34. [[CrossRef](#)]
- Schweizer, Paul. 2019. Triviality arguments reconsidered. *Minds and Machines* 29: 287–308. [[CrossRef](#)]
- Sejnowski, Terrence J. 2023. Large Language Models and the Reverse Turing Test. *Neural Computation* 35: 309–42. [[CrossRef](#)] [[PubMed](#)]
- Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *arXiv* arXiv:1409.3215.
- Van Gulick, Robert. 2001. Reduction, emergence and other recent options on the mind/body problem. A philosophic overview. *Journal of Consciousness Studies* 8: 1–34.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Paper presented at 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, December 4–9; Red Hook: Curran Associates Inc., pp. 6000–10.
- Wang, Xiaoyong, and Junbo Xu. 2023. From the Buddhist Transcendental Epistemology to Viewing the Limitations of AI and a Recognition of Ontology. *Computer Science Mathematics Forum* 8: 37.
- Weise, Karen, and Cade Metz. 2023. When A.I. Chatbots Hallucinate. *New York Times*, May 1.
- Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Translated by Elizabeth Anscombe. Oxford: Basil Blackwell.
- Yarowsky, David. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. Paper presented at 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA, USA, June 26–30; Cambridge: Association for Computational Linguistics, pp. 189–96.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.