Briefly Speaking

COUNCIL ON HIGHER EDUCATION

Number 36
December 2025

Artificial Intelligence and the Sanctity of Science: A Call for Human-Centred Innovation in Higher Education

Abstract

This piece of *Briefly Speaking* interrogates the sanctity of science—its epistemic merit—as a distinctly human endeavour in the context of the growing integration of artificial intelligence (AI) within higher education. While AI offers opportunities for efficiency and innovation in research, teaching, and administration, its uptake often reflects a technocratic logic that risks displacing the ethical, interpretive, and social dimensions that have long grounded the scientific enterprise. By examining the epistemic debates and stakes *vis-à-vis* AI within higher education, this piece of *Briefly Speaking* argues that higher education institutions are at an epistemic crossroads: either to reassert the humanistic foundations of science or to acquiesce to a mechanised model of knowledge production. Making an epistemic argument, the contention is that preserving the sanctity of science requires resisting the reduction of knowledge to data, and reduction of education to algorithmic optimisation, owing to the inability of AI to be epistemically reflexive. It calls for an intentional reaffirmation of science as a human pursuit—rooted in curiosity, moral responsibility, and dialogic engagement—especially at a time when AI threatens to obscure these foundations in favour of scale, speed, and predictability.

Keywords: Science, artificial intelligence, knowledge, philosophy, human endeavour

Introduction

While there has been a proliferation of scholarship considering the impact of artificial intelligence (AI) and what is termed ‘existential risks (x-risks) from AI’ (Kasirzadeh, 2025), this body of work does not give us much insight into what is significantly essential about the human factor in knowledge creation. Briefly, x-risks arising from AI can be categorised into two schools of thought. The dominant thinking has centred on ‘decisive risk’, wherein AI would harm human beings in one cataclysmic event, a scenario depicted in the paperclip maximiser thought experiment developed in Boström’s (2003) work. Kasirzadeh (2025: 1981) argues for a more nuanced perspective, suggesting instead that we pay attention to the ‘accumulative AI risk hypothesis’, whose outline fits best the context of higher education and the implication of AI on knowledge production: "AI x-risk results from the build-up of a series of smaller, lower-severity AI-induced disruptions over time, collectively and gradually weakening systemic resilience until a triggering event causes unrecoverable collapse."

Demonstrating the point and drawing on Ord's (2020: 28) thinking, Kasirzadeh (2025: 1982) outlines how the two schools differ by detailing that:

> Nuclear weapons and climate change have striking similarities and contrasts. They both threaten humanity due to major shifts in the Earth's temperature, but in opposite directions. One burst in upon the scene as the product of an unpredictable scientific breakthrough; the other is the continuation of centuries-long scaling-up of old technologies. One poses a small risk of sudden and precipitous catastrophe; the other is a gradual, continuous process, with a delayed onset-where some level of catastrophe is assured and the major uncertainty lies in just how bad it will be.

Higher education and the accumulative AI risk hypothesis

In higher education, it would be challenging to argue for a catastrophic event that would render researchers and administrators redundant, given that the project of teaching and learning in Africa remains highly defined by unequal access to and deployment of AI. Kasirzadeh's analysis is therefore helpful in examining the gradual effects that AI has on the enterprise of knowledge. The gradual hypothesis reveals how AI is displacing the human factor in expertise, suggesting that overreliance on AI has at least one significant implication. The technology is often used without effective human oversight, particularly when employed by students who are still developing the requisite skills in their discipline, whether theoretical physics, philosophy, sociology, or economic history.

The thinking, here, points to the reality that the student of sociology, philosophy and economic history is not yet fully aware of the disciplinary methodology that is to be deployed in thinking about the questions that are considered in the discipline. This is more evident in the context of the student who is still undergoing formative training in the discipline. This is not to take away from the agency of the student, insofar as Kumalo (2018) has previously argued that the presupposition that the student is without prior knowledge, before entering the university, perpetuates the problems of creating 'natives of nowhere'. However, it acknowledges that the student is still acquiring knowledge, which means their use of AI might differs from how someone with advanced disciplinary training might deploy and utilise it.

Tarsney (2024) clarifies the point further when he argues that the effect we are most concerned with is not truthfulness or falsity, but rather its consequences. To be clear, in being apprehensive about student use of AI in their work, we must move away from the categories of truth and rather look at what are the implications of their use of AI on the knowledge production project, as individuals who are still being trained in the disciplines. Examining deception derived from AI information as manipulation which misleads, Tarsney argues (2024: 1873) that, "We might summarise this characterisation by saying that it treats deception and manipulation as forms of *misleadingness*, understanding misleadingness in terms of leading an addressee away from the beliefs and choices she would endorse under semi-ideal conditions". It is important to note that without adequate disciplinary training, students may be far more susceptible to being misled by AI than those who are well-versed and trained in the discipline. In other words, the probability of those who are on a learning programme in the discipline being misled is more apparent when we confront the reality that they do not *yet* possess the epistemic reflexivity that Kasirzadeh argues is necessary for us to classify knowledge—specifically scientific discovery—as epistemically meritorious.

It is a serious indictment to higher education that the excitement about AI has resulted in overlooking the concerns about the epistemic stakes of developing knowledge claims. This has sadly gone underexamined in the literature on AI use in knowledge production. This is not to suggest that there has been no work done on the function of tools that aid in knowledge production[1], but instead seeks to clarify that in the wake of contemporary tools like GPT5, Claude 4.5, DeepSeek—V3 and other large language models (LLMs), the research that preceded the introduction of ChatGPT into higher education—and tools like it, globally—and the epistemic stakes of knowledge development have been largely neglected. A consideration that might be useful to contextualise the discussion concerns the question: what is lost in knowledge development when it is taken over by machine learning, computational algorithmic optimisation, and LLMs? The aim of this piece of *Briefly Speaking*, subsequently, is not to defend a position on whether AI should inform knowledge development. Instead, it aims to outline the ethical implications of such developments in knowledge production.

At the core of the argument is a philosophical problem that can be articulated as follows: What makes scientific inquiry a sacred or sanctified human endeavour—rooted in moral responsibility, epistemic humility, and social dialogue—and how might the integration of AI into higher education, even when ethically regulated, risk displacing these humanistic foundations by privileging reductive, technocratic imperatives over the deeper purposes of the university?

Epistemic stakes and their defensibility

In response to the above question(s), epistemic considerations emerge with much-needed attention. Knowledge is knowledge insofar as a claim is verifiable by and through epistemic veracity (*cf*. Fallis, 2000) and scientific methods. To demonstrate the importance of the ethical underpinnings of the epistemic force in scientific discovery, consider that human beings can and

[1] For a more sustained discussion on the question of AI and epistemic tools that aid in the development of knowledge development see Alvarado (2023). His discussion centres on the role of AI as an epistemic technology and he is at pains to distinguish arbitrary ascriptions of the concept 'epistemic technology' to any and all forms of tools that assist in knowledge development. Alverado wishes to ensure that the tools classified as adding value to the knowledge development enterprise are precisely so defined and understood.

are held accountable for the conclusions that they come to in scientific inquiry. While academic freedom, which led to the development of the Manhattan Project that produced the atomic bomb can be used as an objection to the notion of responsibility, this objection can be addressed simply. While those who make scientific discoveries that undermine the survival of the human species, through aid in the project of human annihilation, a discovery like the atomic bomb cannot be equated with its deployment in wiping out humanity. Simply, this scientific discovery cannot be equated with the x-risk AI factors discussed above, insofar as the bomb cannot self-detonate due to the reasoning that human beings pose a problem to the continued sustainability of the planet.

To analyse the problem of humanistic accountability more clearly in scientific discovery, consider that the decisions we make about scientific discovery are different from the discovery itself. While the author can be accused of reductionism for this example, it does demonstrate the problem with clarity. A knife can be conceptually framed as a tool for kitchen use, while it can also be used to inflict harm on another human being. The killer who uses the knife to murder someone cannot claim that the knife made them do it, lest they are suffering from schizophrenic dissociation.[2] Further still, the knife cannot be held accountable for its use in killing a human being. This is how the author addresses the objection to the atomic discovery, insofar as the same discoveries can be applied to nuclear pebble bed reactors for energy production rather than automated weapons deployment. Accountability, in this instance, refers to what Kasirzadeh calls the ability to integrate knowledge into an existing epistemic grid that would account for the discovery to be classified as such. It is the discoverer—acting in consort with their epistemic community—who has the agency to determine how the discovery is to be used and how it fits into the existing disciplinary epistemic grid. An AI, conversely, does not have the reflexive capacity, yet, to be held accountable for the information it provides to its addressee or how that information is deployed.

Tarsney (2025) suggests that accountability can be attributed to the developer, which highlights the complexity of AI in the context of knowledge development projects. The developer can develop an AI whose sole function is to produce misleading information. However, Tarsney's (2025) analysis itself does not concede the point that agential AI, which intentionally produces false information of its own volition, is a valid concern. Instead, his argument demonstrates the reality that misinformation is developed in relation to the prompts that an LLM is fed and prompted to generate. Tarsney's argument is persuasive in the context that his discussion unfolds with respect to political information, its uses and weaponisation against political opponents or persuasion, either in

[2] Even under such circumstances, there are conditions that would have to be met sufficiently, in order for a court of law to come to the decision that, indeed, the murderer suffered from some medical problem. This is to demonstrate that it is not sufficient to suggest that the knife told me to kill someone.

promoting a particular political candidate or driving the addressee to a different conclusion from what they would otherwise come to, given what he terms (Tarsney, 2025) 'semi-ideal conditions'.

To return to the consideration of epistemic veracity, then, veracity refers to the truthfulness, reliability, and integrity of knowledge claims – the extent to which a piece of knowledge or information can be considered accurate, credible and legitimately produced within an epistemic (knowledge—producing) system. Epistemic veracity in this case suggests that the information given to the addressee is neither misleading in the Tarsney (2025) sense nor inaccurate, allowing the addressee to draw concrete conclusions based on its reliability.

But what is epistemic veracity? It might be helpful to define it and its deployment here. To begin, four considerations are necessary in determining the veracity of a claim, and they are: (a) truthfulness, (b) justification, (c) authority, and finally (d) epistemic responsibility. Epistemic responsibility denotes that the developer can be held accountable for their scientific knowledge, LLM programming, and information disseminated for human use.[3] Explicitly, these four considerations give credence to their importance in determining epistemic veracity. Under truthfulness, we ought to inquire whether the information is accurate and reliable. We ask this question because AI can produce confident but false statements (what is termed in the literature, 'hallucinations'). Under justification, we must ask whether there are grounds or evidence to support what is claimed, and this is important because academic knowledge requires traceability, including clear sources, sound reasoning, and well-defined methods. Traceability becomes even more critical when the reader considers how information is disseminated, leading to congressional—or other law-making bodies—mandating reporting, as the said information has had implications for social well-being, as detailed in the footnote below.

When considering the authority of the claims made, it is essential to critically reflect on who or what the knower is. A non-human 'generator' disrupts traditional authorship conventions and accountability, as detailed above. Finally, considering epistemic responsibility, which derives from authority, we need to ask who bears the duty of information verification. Students and academics alike risk outsourcing judgments to AI, which does not have an ethical framework that

[3] Consider that in the wake of the effects of the vaccines that were developed by Moderna, the company—through its Chief Executive Officer and senior executives—was discovered to have deleterious effects on humans. An increase in myocarditis datasets, among young men because of the higher dose of the Moderna vaccine, led to direct Senate scrutiny in the United States, demonstrating the point of human scientific discovery being held accountable, which cannot be said for AI, as it does not have the same safe proofs that human scientific systems are thought to have.

can accurately determine what is good or bad, as demonstrated by D'Alessandro's (2024) exposition of 'Deontology and safe artificial intelligence'. Moreover, in this move to outsource information verification, the AI cannot be held accountable for the information and how it is used. As demonstrated in the preceding discussion, the human agent that prompts the AI is the actor responsible for developing the information, its uses, and deployment, and can thus be held accountable for that information. In the case of the human actor as discoverer, can it be said that the AI is partly responsible for a scientific discovery when it has been prompted to arrive at a scientifically and epistemically meritorious one? This is discussed in more detail in the ensuing section, however, when the ethical implications of the argument is discussed.

Why veracity is not a technical, but an epistemic issue can be summed up as a correlational derivative between epistemic authority, labour, accountability and justice—making the concerns of AI interrelated to the calls for decolonisation as they are interested in the attainment of epistemic justice. To be sure, AI cannot be said to be a knower; it is a statistical synthesiser, which means that assigning epistemic weight to its outputs shifts authority from *thinking selves* to systems and algorithms. This shift sidesteps an essential aspect of accountability, which, as detailed, is necessary when considering how information is not only used but taken up and applied in the scientific community and society more broadly. Moreover, epistemic labour denotes the processes of reasoning. Overreliance on AI externalises this labour while undermining the cultivation of disciplinary methodology in graduate and undergraduate training. Finally, epistemic justice—or injustice—can be attributed to the biases LLMs are inherently encoded with. Insofar as LLMs are developed in the northern epistemically dominant Euro-American context, the knowledge of much of the world (African, Indigenous, marginalised communities) would continue to go underrepresented. The result of this underrepresentation is the skewing of veracity toward Euro-American epistemic canons, which decolonial scholars are calling for pluralisation and democratisation, *à la* Kumalo (2021).

Alvin Goldman's (1999) concept of veritistic social epistemology (VSE), which was first detailed extensively in his *Knowledge in a Social World,* becomes essential in demonstrating that work in this area has long been developing, even as it may be superseded by more recent scholarship. Goldman's thesis claim can be summed up as an evaluation of our "[…] practices by their causal contributions to knowledge or error". In simple terms, the field advanced by Goldman and other veritistic epistemologists "evaluate[s] social practices with regard to their impact on knowledge acquisition" (Fallis, 2000: 305). The impact of using and deploying LLMs in higher education has the implication that students may fail to cultivate adequate skills, which can hinder their ability to be regarded as competent in the area without sufficient knowledge in the relevant field. This has serious ethical implications for the institutional certification of students training in disciplines such as medicine, accounting, education, and law. Once we consider the four quadrants detailed above—truthfulness, justification, authority, and epistemic responsibility—it becomes essential to determine how AI is impacting the practices of knowledge production, insofar as it is a human endeavour.

In doing so, the argument developed here aligns with Ganapini's (2025: 1) thinking that "[…] AI can find new scientific facts based on some background theory", while distinguishing this reality from the fact that scientific discovery (*ibid.*), which can be defined as "an epistemic achievement that requires a level of competence and self-reflexivity that AI does not yet possess". To expand on the first claim, Ganapini (2025: 3) writes "[…] because making a scientific discovery is an epistemic achievement – the 'agent' discovering needs to be someone able to assess her own competence and integrate that information into her reasoning". What goes unstated in her argument is the implicit assumption of methodological ability, which means that the discovering agent needs to be aware of the conventions that govern a particular area of knowledge. Without this awareness, the discoverer can be said to have stumbled upon information without the necessary driving force of seeking to advance scientific understanding.

Insofar as one can stumble upon information without adequate means to discern its significance, this piece of *Briefly Speaking* now proceeds to consider the ethical implications of training and using artificial intelligence in the training process in the ensuing section.

The moral argument

Two examples are used to clearly articulate the ethical implications of using AI in knowledge development. First, consider the reality of visiting a hospital and being examined by a doctor who has qualified using AI in their medical training. While the scenario is laughable, the ethical implications are major. To begin with, the doctor's judgment would be in question regarding their ability to make informed medical decisions due to knowledge gaps about human physiology. AI in medical treatment, diagnosis and examination is said to be an epistemic enhancer in the way that Humphreys (2004) defines it. Humphrey's concept of an epistemic enhancer is challenged, however, by Alvarado's (2023) analysis on the premise that the concept is too broad and can lead to any enhancers that allow us to do science being classified as epistemic enhancers. Alvarado (2023: 2) argues that "We must […] distinguish what something is designed, developed and deployed for from what something does and how it does it". Moreover, misdiagnosis and mistreatment would not only expose the practitioner to ethical and legal liability but would also call into question the institution where they were trained.

The use of AI in medical training, then, has at least two implications. First, legal liability on the part of the practitioner, their employer, be it a private medical facility or a public facility. It undermines the public's confidence in medical practitioners in the second instance. Simply, patients seeking medical treatment would be cautious about where and who is giving them treatment, based on a lack of confidence in the abilities of practitioners due to their reliance on AI. This second implication points to a deeper consideration, namely that while we might rely on AI for the efficiency of treating patients and drawing diagnostic conclusions, the importance of the individual being treated comes to the fore with more pressing concerns. Simply, the human focus highlights the ethics of overreliance on AI, as the human is being treated with undermined ethical considerations, particularly when the medical practitioner relies on AI in treating the patient.

This should not be understood as completely dismissing the importance and necessity of AI and technology in advancing and enhancing patient medical treatment. Alvarado (2023: 2) suggests that there have been several epistemic tools, which he defines as tools "[…] explicitly deployed in such contexts to manipulate epistemic content such as data, *and* it [an epistemic tool] manipulates such content specifically through epistemic operations such as inferences, predictions or analysis". His definition of epistemic tools highlights a key component in Kasirzadeh's (2025) definition that disqualifies AI from making epistemic claims insofar as the epistemic tool can only come to the said conclusion on the premise that the discoverer is able to understand the epistemic content in an ecosystem—epistemic grid—of the field in which they are operating.

The ethical implications of AI use in this example underscore the importance of the centrality of human beings who use the tools to enhance their understanding. AI does not yet possess the reflexivity necessary to draw epistemically meritorious conclusions. When the said conclusions involve how human beings are treated, it suggests that there needs to be ethical guidelines with respect to how AI is used, while displaying the centrality of the question posed in the introduction concerning the sacredness of science as a human endeavour. The lack of autonomy with respect to AI corroborates the conclusions that Kasirzadeh (2025) draws in their argument.

While clear in the case of biomedical science, it is important to consider the ethical implications of AI in the humanities and the arts. To demonstrate this point, the reader's attention is drawn towards the challenge posed by Andy Warhol in the advent of pop culture. The question that Warhol makes apparent is the function of the human element in the development of art, which is to say that which is sacred about the human in artistic creativity.

As this argument has established, AI cannot be said to have the capacity to generate epistemically meritorious claims, but can it be said to have the capacity to generate artistically inspired works? The assessment of the aesthetic object generated by AI should not be held to the stringent requirements detailed in Kant's *Critique of Judgement* (1790), in answering the question posed. However, this work is brought up as a method of demonstrating the ethical implications of aesthetic objects, which is to say that it is not without its historical location and merit in scholarship.

When examining the historical traditions that have defined aesthetic debates, one can assume that the rise of Andy Warhol, and his subversion of art theory and practice in his creation of the genre of pop art, can be read along the lines of AI and its uses in developing aesthetic objects. The use of popular media, massified replications, and a non-human style can be seen as taking away from the artistic genius that defined the traditions that had come before Warhol. In the description used above, involving popular media, massified replications, and a non-human style, can also be read as a descriptor of what we find in the realm of artworks that are developed by AI. In this instance, can the argument be made that

Warhol's innovations are the substantiating grounds for the claim that AI can come up with aesthetically pleasing works to the human eye? Does the human creator, in creating an aesthetic object, have something sacredly human about the project or the process of creativity?

These questions point to the reality that there is something sacred that we regard as reservedly human that defines what is aesthetically, or at least, what allows us to come to aesthetic judgments that can be classified or that claim to exist in the domain of the sublime—in the Kantian sense. Moreover, when examining how AI produces artwork, there is a mechanistic sense—the machine produces the work, thus the work ceases to fall in the classification of art. That is, it is not art, insofar as it is predetermined by a machine using the machine's algorithmic synthesis of what is known to the machine about the different traditions and epochal definitions of art. This is indicative of the fact that the artistic traditions that are undertheorised, less known, because of the erasure and epistemic slighting that often befalls the inventions and innovations coming from the African continent, as but one example, cannot influence and contribute to the generation of the aesthetic object. And yet, a human being – situated in the Democratic Republic of Congo, for example – trained in Classicism, Renaissance art, the Baroque movement, Neoclassicism, Romanticism, Realism, Impressionism, Symbolism, Fauvism, Expressionism, Cubism, and Futurism can combine this knowledge with their own methods to generate a new tradition that would define the artworld, as we know it.

Simply, the machine does not have the ability to generate new areas of innovation, even within the realm of the aesthetic, yet. This conclusion comes from the fact that the machine is unable, in the sense Kasirzadeh (2025) writes about, to integrate knowledge the way an agential human being can. That is, LLMs are only able to synthesise information on the premise of the prompts that are fed into them by human actors, who are in search of certain outcomes or conclusions.

Conclusion

This piece of *Briefly Speaking* has traced the argument using the simple question, 'What is so sacred about science and knowledge that forestalls our accrediting artificial intelligence with the ability to create knowledge'? A persuasive answer, that is not yet conclusive, points out that AI is unable to arrive at epistemically meritorious discoveries of its own volition. Simply, AI needs human agency, which prompts an LLM to generate certain answers that are only understood by the human requesting said answers in the context of a wider epistemic awareness that the human brings to the questions that they pose to the AI. In this sense, there is a distinction being made between a conclusive *versus* a persuasive argument. While the reader might be persuaded by the argument presented in this piece of *Briefly Speaking*, this does not foreclose the ability to come to conclusive answers with the development of more research in this area.

This is simply to say that higher education—globally—will need to clearly define what is so important (what is termed the sacred in this argument) about scientific knowledge and how it is that AI can either encroach or be limited from encroaching on the scientific enterprise. Without a clear understanding of the epistemic and ethical stakes concerning this primordial question, that is to say that without the sanctity and sacredness of scientific knowledge discovery, higher education will continue to equivocate on the uses and limitations of AI in the teaching, learning and research enterprise of the university, more generally.

References

Alvarado, R. 2023. AI as an epistemic technology. *Science and Engineering Ethics*, 29(5): 1—32.

Boström, N. 2003. Are we living in a computer simulation? *Philosophical Quarterly*, 53(211): 243–255.

D'Alessandro, W. 2024. Deontology and safe artificial intelligence. *Philosophical Studies.* 182: 1681—1704.

Fallis, D. 2000. Veritistic social epistemology and information science. *Social Epistemology*. 14(4): 305—316.

Ganapini, M. B. 2025. *Can AI make scientific discoveries*. Philosophical Studies (submitted).

Goldman, A. 1999. Knowledge in a Social World. Oxford: Oxford University Press.

Humphreys, P. 2004. *Extending Ourselves: Computational Science, Empiricism and Scientific Method*. Oxford: Oxford University Press.

Kasirzadeh, A. 2025. Two types of AI existential risk: Decisive and accumulative. *Philosophical Studies*. 182: 1975—2003.

Kumalo, S. H. 2021. *Decolonisation as Democratisation: Global Insights into the South African Experience*. Cape Town: HSRC Press.

Kumalo, S. H. 2018. Explicating abjection – Historically white universities creating natives of nowhere? *Critical Studies in Teaching and Learning*, 6(1): 1—17.

Ord, T. 2020. *The Precipice: Existential Risk and the Future of Humanity*. London: Bloomsbury Publishing.

Tarsney, C. 2025. Deception and manipulation in Generative AI. *Philosophical Studies*, 182: 1865—1887.

Disclaimer and acknowledgements

It is acknowledged with sincere gratitude that this specific issue of *Briefly Speaking* was researched and drafted by Prof Siseko H. Kumalo. The CHE also expresses its profound appreciation to the peer reviewers who contributed immensely to shaping this specific issue of *Briefly Speaking*.

Comments should be addressed to research@che.ac.za.