The Future of Philosophy: How to do Things with Artificial Intelligence

Ioannis Votsis

1. Introduction

Philosophy is one of the oldest subjects. It has graced the curricula of academic establishments since the birth of modern universities in the eleventh century. Antedating those establishments, it has been a leading subject in the higher learning schools of antiquity, the most notable examples of which are Plato's Academy and Aristotle's Lyceum. It is a subject that has evolved, sometimes breaking apart and forming new disciplines, e.g. Physics growing out of Natural Philosophy, and at other times (partially) merging with other disciplines to form new branches or schools, e.g. medieval Scholastic Philosophy, combining Aristotelian Philosophy and Christian theology. More recent turning points, include the development of Formal Logic, the Analytic vs Continental divide, the Linguistic turn, Experimental Philosophy (X-Phi) and Formal Philosophy. Despite the ebb and flow of its content and methods, one thing remains constant. Philosophy is principally concerned with answering foundational questions. It tends to adopt normative and meta-theoretical stances. Moreover, it tends to employ, more readily than other disciplines, certain methodologies, e.g. dialogues, thought experiments, and conceptual analysis.

Despite its longevity, Philosophy is an academic subject in peril. One key threat arises from short-sighted, often commercially driven, interests that put pressure on universities to curtail or close programmes or departments that are deemed unprofitable or not worth the effort. Yet, this is not the only, or indeed the gravest, threat. Artificial Intelligence (AI) promises to revolutionise everything we do, including the way we teach, learn, and research subjects. In this paper, we explore some threats and opportunities afforded to Philosophy by AI and related computational techniques, focusing on the opportunities, particularly how those technologies can help boost and transform some of our most cherished methodological frameworks. By way of illustration, we offer some concrete suggestions of how to do things with AI and computation, highlighting advantages along the way.

The paper is structured as follows. Section 2 provides a brief overview of current Philosophy as an academic subject. The section is split into two sub-sections. The first circumscribes Philosophy via the content of its investigations. The second circumscribes it via the methodological frameworks it relies on to carry out those investigations. Section 3 briefly identifies some threats to Philosophy as an academic subject. Section 4, the most substantial of the lot, presents opportunities afforded to Philosophy by AI and computation. Its focus is on how AI can amplify and revamp Philosophy's methodological frameworks. Section 5 offers some speculative thoughts about the future of Philosophy, particularly on which areas of Philosophy are likely to adopt those amplified and revamped methodological frameworks and why. Section 6 summarises the main points and leaves the reader with some parting thoughts about what our responsibilities are towards young Philosophers facing an uncertain future.

2. What is Philosophy?

In what follows, we provide a rather cursory review of Philosophy based on its content and the methodological frameworks it regularly employs. We treat each of these in distinct sub-sections. Before we swing into action, a couple of qualifications are in order. First, our review is limited in at least two ways. First, it is limited to western Philosophy. This obviously leaves out several other approaches, including African, Asian and Indian Philosophies. Second, it is limited to the subject as it is primarily practiced in an academic setting. This obviously leaves out Philosophy as it is practiced in film, literature, and the various arts. No claim is made here that what is excluded lacks value. Rather, the imposed limitations are primarily personal, i.e. they reflect our own familiarity with the subject qua academics in the western Philosophical tradition.

2.1 Content

There are several approaches to circumscribe Philosophy through its content. One approach is by looking at how the discipline is actually structured. That structure is not set in stone, but it is fairly well drawn, motivated, and understood, to ignore. Barring additions to the periphery, it has also remained largely unchanged for centuries, arguably going back to antiquity. Roughly speaking, we can divide its content into theoretical-leaning and practical-leaning. Many universities in continental Europe continue to draw this distinction to this very day, and even have separate theoretical and practical Philosophy units, each with its own dedicated chair, but inhabiting the same departments or institutes. Roughly speaking, Practical Philosophy is an umbrella term under which fall the areas of Ethics, Political Philosophy and Aesthetics. Sometimes these are also referred to as Value areas. Equally roughly, Theoretical Philosophy is an umbrella term that covers Metaphysics, Epistemology and Logic. These six areas are often referred to as the 'core' areas, or branches, of Philosophy.

Many other sizable areas exist, notably the Philosophies of Art, Language, Logic, Mathematics, Mind, Religion, and Science.[1] Let us call these sizable areas 'primary' for convenience. Primary areas still raise the same questions as the core areas, but they do so in a contextualised manner. Philosophy of Science, for example, still asks what exists (a core metaphysical question), what we can know about what exists (a core epistemological question), and how we can/ought to reason about what exists (a core logical question) but it restricts such questions to the scientific context. To illustrate, it asks questions such as 'What are laws of nature?', 'What can be known about the world through our best scientific theories?' and 'Whether second-order evidence, i.e. evidence about evidence, should be employed to support existence claims made by scientific theories?'.

In recent years, the number of areas in Philosophy has ballooned to include Computational Philosophy, Meta-Philosophy and the Philosophy of AI, occupying centre stage in the current paper, as well as the Philosophies of Action, Biology, Chemistry, Economics, Education, Film, History, Information, Law, Medicine, Sports, Physics and War, to name but a few. For convenience, let us call these 'secondary' areas. Some of them can be taxonomised under primary areas, e.g. the Philosophies of Biology, Chemistry, Economics, Medicine and Physics fall under the Philosophy of Science. Others stand alone or at least possess more autonomy, like the Philosophies of Action, Education and Law.

The abovementioned secondary areas have grown sufficiently to command serious scholarship. Others, not mentioned here, are still nascent and typically lack sophistication. In fact, these days there is a tendency to prefix any noun, proper name, verb, or adjective with the expression 'Philosophy of', leading to such specialised areas as the Philosophies of the Absurd, Dance, Lying and the Matrix. Cynics see these as disingenuous attempts to carve out areas for scholars to lord over as experts. Perhaps some of them are like that or at least begin life like that. Whatever the motives for their creation, what ultimately matters is whether these areas increase our understanding of the relevant subjects and do so in a distinctively philosophical way.

Another approach to Philosophy's circumscription is through its preoccupations. One major preoccupation is with foundational issues, embodying the search for general principles, concepts or posits, upon which other principles, concepts or posits may rest, be explained by or

[1] To be clear, we are here referring to the 'Philosophies of', therefore the 'Philosophy of Logic', not Logic itself. As we just saw, Logic is characterised as a core area. The Philosophy of Logic is characterised as a primary area.

somehow accounted for.[2] An instructive example here compares a Mathematician to a Philosopher of Mathematics. While a Mathematician asks questions about how to prove the formal properties of objects within a given branch of Mathematics, the Philosopher of Mathematics asks which, if any, proof system is the right one. The sharpness of the distinction between researchers of a field *x* and Philosophers of *x*, as well as of the distinction between foundational and non-foundational issues, should be taken with a pinch of salt. Moreover, there are plenty of researchers who are not formally trained in Philosophy, but who nonetheless have genuine philosophical interests, and even pursue foundational issues in their work. Similarly, there are researchers who are formally trained in Philosophy, but have non-Philosophical interests, and even pursue non-foundational issues in their work.

A related preoccupation, and hence another approach to circumscribing some of Philosophy's content, is through the tendency of its practitioners to adopt a meta-theoretical stance. All sorts of discussions in Philosophy tend to be conducted at a higher level. Taking the discussion a level up, enables us to see the bigger picture, which helps with generalising as well as drawing (dis-/)analogies between lower levels. Examples include Plato's encouragement to look for the real causes 'behind' the shadows, i.e. the appearances, in the allegory of the cave (see Plato's *The Republic* in Cooper 1997), Kant's pursuit of what conditions make metaphysical knowledge possible (see Kant's *Critique of Pure Reason* in Guyer and Wood 1998), and Wittgenstein's examination of the linguistic origins of Philosophical problems (see Wittgenstein 1953).

Meta-theoretical concerns are, of course, not unique to Philosophy. They can be found in all disciplines, including Economics, Physics and Biology. In Economics, (some) researchers take a step back and ponder whether we can cogently model behaviour without taking into account realistic conditions underlying decision-making and reasoning. This attitude has led to the burgeoning sub-field of bounded rationality. In Physics, (some) researchers worry about how to understand the formalisms of quantum mechanics, offering competing interpretations that vie for supremacy. In Biology, (some) researchers stand back and ask how to demarcate the category of species. Even though meta-theoretical concerns permeate all disciplines, Philosophy is uniquely fixated on them. This becomes obvious when one considers that such concerns and those who entertain them are more likely to be found on the margins of those other disciplines. Unsurprisingly, those same individuals are more likely to present or publish their works in Philosophy conferences, books or journals.

Yet another approach to circumscribing Philosophy's content is through its preoccupation with normative considerations. Many disciplines, particularly the empirical ones, ask questions about how things are, not how they ought to be. Since Philosophy is, by and large, not an empirical discipline (for an exception, see the discussion on X-Phi below), it cannot but divert its attention to normative questions. Examples include the questions: How ought we to live? What might the proper standards of knowing be? How should beauty be evaluated? Moreover, normative questions offer opportunities for meliorative work and Philosophy is particularly enamoured with such work. For example, Philosophers of Science routinely employ rational reconstruction to show what ought to have happened in a scientific revolution, as opposed to what actually did happen, often devising general recipes for scientific theory change in the process. As with meta-theoretical concerns, it would be foolish to claim that normative considerations are unique to Philosophy. Still, Philosophy is arguably at least as equally fixated with such considerations as any other normative-heavy discipline.

[2] This is not the same thing as reductionism, though the two notions are clearly related. For example, to be concerned with foundational issues means to ask questions such as *whether* one set of posits reduces to another set of posits. That question can be answered both in the affirmative and in the negative. Only the former leads to reductionism.

The above approaches to circumscribing Philosophy's content are not faultless. They should be read as strong tendencies, especially relative to (most) other disciplines. Moreover, there are other approaches that were left out. One example is circumscription through Philosophy's emphasis on interpretative work. This emphasis has even spawned an entire methodological framework called 'Hermeneutics'. Alas, there is not enough space to discuss this and other approaches here. What we will say, however, is that, much like Hermeneutics, the meta-theoretical and normative stances do not just affect the content of Philosophy. They are first and foremost methodological frameworks. It is to the circumscription of Philosophy through its methodological frameworks that we now turn to, as they are the main focus of this paper.

2.2 Methodological Frameworks

In this sub-section, we focus on a collection of methodological frameworks employed in Philosophy, namely dialogue, thought experiments, conceptual analysis, meaning as use, explication, reflective equilibrium, and X-Phi. These frameworks have been chosen because they are both popular and ripe for a computational / AI makeover.

Before we proceed, a number of provisos are in order. First, though fairly abstract, these methodological frameworks are presumably more concrete than the meta-theoretical and normative stances discussed in the previous sub-section. For example, after adopting a meta-theoretical stance, there is still plenty of flexibility regarding the way one goes about arguing for a specific claim. The listed frameworks can help fill that gap. Second, these frameworks need not be in competition and may even sometimes be deployed jointly. One example is the use of thought experiments and conceptual analysis in attempting to settle the meaning of Philosophically important concepts. Third, the list is not exhaustive, by any means, but it provides a cross-section of regularly employed methodological frameworks in Philosophy. Here are some other methodological frameworks, besides the meta-theoretical and normative stances as well as hermeneutics, that we left out: dialectics, historical materialism, methodological individualism, methodological holism, phenomenology, positivism, pragmatism, rational reconstruction, social constructivism, structuralism, transcendentalism, and verificationism. Fourth, two important omissions from our list are logic and formal philosophy. Logic is omitted primarily because attempts at its computationalisation, including its integration with AI, have been around for a long time. The first successful AI program is arguably *The Logic Theorist* (Newell and Simon 1956), able to prove several theorems from Whitehead and Russell's *Principia Mathematica*.[3] There is thus no need to discuss this framework in Section 4, and, hence, no need to bring it up here. Formal Philosophy is omitted primarily because it is already open to the employment of computational methods, seeing as it counsels the use of any and all formal tools to solve Philosophical problems.[4] Fifth, some of these frameworks are also applicable, and, in fact, applied, either explicitly or implicitly, in other disciplines. Thought experiments, for example, are frequently found in Physics and beyond. Sixth, as with content, the circumscription of Philosophy via methodological frameworks is imperfect. These frameworks merely reflect tendencies within the discipline or segments thereof.

[3] Since then the history of logic and AI have been intertwined, with Philosophers making significant contributions. For example, Hilary Putnam and John Alan Robinson are key figures in the development of efficient automated theorem proving methods, notably through their work on resolution and unification respectively. More recently, John Pollock's work on defeasible reasoning has been at the core of Computational Argumentation methods.

[4] The highly successful Munich Centre for Mathematical Philosophy, founded by Hannes Leitgeb, is devoted to formal philosophy and has been mass producing quality outputs since the early 2010s.

Let us begin with dialogue. Utilising dialogue, whether real or imagined, has been a favourite methodological framework of Philosophers probably since the subject's dawn. Plato's dialogues are a paradigm example.[5] A proposal is made by an interlocutor, e.g. love is a longing to reunite with one's other half, and discussion ensues. The discussion tends to be rather formulaic. In one commonly employed formula, the proposal quickly runs into trouble, i.e. objections, sequentially forcing adjustments that ultimately leading to a significantly revised proposal or its abandonment. In another, adherence to the proposal remains steadfast, as each and every attempt to counter it leads to contradiction. Often, one interlocutor leads the discussion and is likely to be the one dispensing wisdom or dispelling wrongheadedness. Dialogues continue to be used by Philosophers to this day, though they are patently more infrequent, and, perhaps, tend to appear in popularised, rather than academic, works. Notable examples of modern-day dialogues include *Dialogues in Limbo* (Santayana 1925), *A Dialogue on Personal Identity and Immortality* (Perry 1978), and *Tetralogue*: *I'm Right, You're Wrong* (Williamson 2015).

The next methodological framework deserving our attention is thought experiments. We may describe these as 'devices of the imagination' that seek to convince the reader of some claim through a priori reasoning, often by stoking our intuitions (Stuart, Fehige and Brown 2017). As with dialogue, a thought experiment is set up to demonstrate that either something or its opposite is the case. Lucretius famously tried to prove that space is infinite via a thought experiment involving spear throwing. Similarly, Aristotle tried to prove that vacuums are impossible – in his own words, 'nature abhors a vacuum'. Thought experiments continue to be used in Philosophy to this very day. Indeed, that use is even wading into debates over AI, sometimes inadvertently. An example is Searle's Chinese Room Argument. An example of one that started out life as a pure moral problem but has found new life in discussions of AI autonomy is the trolley problem.

Consider next conceptual analysis. This framework mimics, admittedly in limited respects, the empirical method of science. The latter proceeds by testing rival scientific hypotheses regarding some target system against empirical evidence about that system. The former proceeds by testing rival philosophical hypotheses regarding concepts against intuitions about what are judicious and injudicious uses of those concepts (Jackson 1998). Philosophical hypotheses regarding concepts take various forms. At the limit, they are formal definitions specified in terms of necessary and sufficient conditions. The key idea here is that the concept being analysed is de-compositionally understood by reference to more basic (potentially un-analysable) concepts. For example, the concept of lying may be analysed in terms of the more basic concepts of asserting a proposition and disbelieving it: Person x *lies* to person y if and only if (i) x *asserts* proposition P to y, and (ii) x *does not believe* P. Analyses are deemed correct provided they fit all relevant intuitions we throw at them. One counterexample is enough to lead to their undoing. Counterexamples are typically not difficult to find. Continuing with the above analysis of lying, one intuition may be that there are plausible cases where both conditions i and ii are met but where x is not lying, e.g. x does not intend to deceive y.[6] If we accept that intuition, the two conditions are deemed insufficient. Another intuition may be that there are plausible cases where either of the two conditions is not met but x is lying, e.g. when x does not assert P but allows P to be implied. If we accept that intuition, we deem one of the two conditions (in this case i) as unnecessary. Either way, via insufficiency or non-necessity, the proposed definition and corresponding hypothesis about the concept are rejected.

The next four methodological frameworks, meaning as use, explication, reflective equilibrium and X-Phi, can be thought of as being informed or motivated by the failures of conceptual

[5] One of the referees commented that dialogues are 'the single most important [method] of a philosopher. The Socratic Method is THE philosophical method, even if we don't write in dialogues anymore'.

[6] The same referee helpfully provided the concrete example of an actor making assertions on stage.

analysis.[7] We begin with meaning as use (Wittgenstein 1953). At the most basic level, this is the idea that the meaning of terms or concepts is solely determined by how they are being used. Since language use is messy, advocates of the meaning as use framework reject the view that concepts can be given proper definitions, i.e. articulated in terms of necessary and sufficient conditions. Instead, they see concepts as fuzzy, connecting their tokens via family resemblance (roughly: similarity) relations. Take the concept bird. The concepts robin, sparrow, heron and ostrich are clearly tokens of the concept bird, even though it is difficult – some even say impossible – to list conditions that all tokens of birds (past, present and future) and only tokens of birds satisfy. For example, not all birds fly. Robins, sparrows and herons do, but ostriches do not. Instead of proper definitions, meaning as use advocates urge us to simply track family resemblance relations between concepts. These establish a graded type of membership, e.g. sparrows are more like robins than ostriches, and, ultimately, fuzzy concepts. It is worth noting that the meaning as use approach has inspired influential cognitive science theories of concepts. Prototype theory, for example, holds that fuzzy concepts are structured by similarity to prototypes, either concrete tokens deemed most paradigmatic (e.g. robin) or abstract tokens created by statistically averaging all tokens (e.g. fairly small, feathered, flying, beak-bearing).

Explication is another methodological framework that takes as its starting point the messiness of ordinary language. It is the brainchild of Rudolf Carnap. Rather than trying to do full justice to our common-sense intuitions or usage of concepts, Carnap (1950) insisted that we must focus on making them logically tidier and, if possible, scientifically grounded. Thus, unlike meaning as use, one can in principle ignore (some) actual usage when evaluating the meaning of a concept and can even extend that meaning to cover conjectured usage, provided this helps the concept to successfully fulfil some role. Moreover, unlike meaning as use, proper definitions are not only permissible but encouraged. Take the Ptolemaic concept of planet as a wandering star. We may adjust the meaning of that concept so that planets are no longer considered to be stars, thereby contracting its extension by removing the Sun. We may also expand that meaning by adding exoplanets into the concept's extension. We do all this to increase coherence with what we know about the world, e.g. stars differ from planets in that they are spheres of plasma undergoing nuclear fusion, and, generally, more exactitude. As such, explication is a pragmatically guided logical and scientific reconstitution of the meaning of concepts. Crucially, Carnap identified four desiderata for explicated concepts: similarity to the explicandum (the original concept), exactness, simplicity and fruitfulness. We return to these in Section 4.

Reflective equilibrium is yet another methodological framework that feeds off of the failures of conceptual analysis. As with explication, but unlike meaning as use, it does not deem definitions of concepts impossible. Moreover, as with explication, it is ready to jettison or modify some usage or intuitions. Indeed, rather than simply testing philosophical hypotheses against intuitions as conceptual analysis counsels, it treats both hypotheses (or principles) and intuitions (or judgments) as having an in principle equal say in the process. Finding out how best to characterise a concept or indeed what philosophical hypothesis is worth endorsing must, according to this framework, be achieved via an iterative process where we sequentially revise both intuitions and hypotheses until we reach a stable state (Goodman 1965). That is, a state where there is no longer a conflict between the hypothesis and the intuitions and where the hypothesis explains those intuitions. Take the concept of a game (cf. Wittgenstein 1953). Our initial intuitions may be that chess, football, basketball, solitaire and fighting with one's sibling are games. Our initial hypothesis may be that games are rule-governed activities that involve two

[7] There have also been other reactions to the failure of conceptual analysis. Williamson (2000) argues that some, and indeed most, concepts in ordinary language are unanalysable. In his own words: "Most words express indefinable concepts" (p. 100). A far more fruitful approach, according to him, is to take such concepts as primitives. His favourite example is knowledge, which is seen by him as a mental state.

or more people, where only one of them can emerge victorious. This hypothesis clearly doesn't fit solitaire, which is played solo. We may thus revise our hypothesis by dropping the 'two or more people' requirement to be more inclusive of such games, of which there are presumably many. Similarly, we may revise our set of intuitions by dropping the 'fighting with one's siblings' example of a game, as it is presumably too much unlike most examples of games in that it is non-rule-governed activity, assuming, of course, that rule-governed activities are a *sine qua non* for games.

Finally, let us consider X-Phi.[8] Proponents of this methodological framework reject or at least curtail the evidential role of intuitions and conceptual analysis. They counsel behavioural experiments, where the intuitions of subjects are elicited and measured, as the best means of determining how language is being used (Stich 2015). Several such studies have been carried out to date, the results of which seem to show that, at least under certain circumstances, intuitions vary across dimensions such as age, sex and culture. What can be concluded from this? One suggested is that intuitions are evidentially worthless in assessing hypotheses about concepts. Another is that not all intuitions have equal evidential worth, and, therefore, not all of them need to be 'saved' by hypotheses about concepts. An extreme version of this view is that only those intuitions that are the product of Philosophical training have evidential worth. Needless to say, this version readily invites accusations of elitism and arrogance.[9] Yet another suggestion is that although some intuitions are evidentially worthwhile, they do not test a single concept, but rather a plurality of them. Thus, even though we may use the same name, say 'causation', for some collection of phenomena, there may in fact be several related sub-categories and corresponding concepts, say 'causation_1', 'causation_2', etc., each with their own idiosyncrasies and associated intuitions (Cartwright 2004; Votsis 2011).

3. Threats

Having examined how Philosophy can be circumscribed, it is now time to consider some prominent threats to it. As already indicated in the Introduction, the focus of this paper is the opportunities, not the threats, thus this Section is inescapably cursory and merely serves the purpose of giving some context to the broader discussion.

We begin with some generic threats to Philosophy as an academic subject and move on to those originating in AI. Philosophy is under siege in some countries – here we restrict our comments to the UK and the US – where it has traditionally been strong. It is under siege by short-sighted, often commercially driven, interests that put pressure on universities to spend less of their resources on such subjects. The result is the curtailment and even closure of programmes and departments. This is particularly the case in the UK, and to a lesser extent the US. In the UK, joint undergraduate and postgraduate enrolments have fallen from a high of 14,565 in 2018-19 to 13,665 in 2024-25. This may not seem like much of a drop. It is, however, compounded by frozen home student fees for undergraduates (these have remained at roughly £9k since 2012) despite rising costs (£1 worth of goods and services in 2012 costs £1.48 in 2026).[10] Recent closures or substantial cuts to Philosophy departments or programmes include the universities of Kent in 2024, Kingston in 2025, and Hertfordshire in 2026. The US has also seen a drop in Philosophy undergraduate students. In 2013, 9,427 Philosophy degrees were conferred, compared to 7,091

[8] This falls under the broader methodological movement of *Naturalising Philosophy* (Kornblith 2002). Explication can also be thought of as falling under this movement in that it emphasises the scientific grounding of meaning.

[9] Its proponents would, no doubt, claim that we should not conflate elitism and arrogance with expertise.

[10] The enrolment figures are sourced from the Higher Education Statistics Agency and can be found here: https://www.hesa.ac.uk/data-and-analysis/open-data-tools/subject-uptake-dashboard. The cost figures were calculated using the Bank of England's inflation calculator (April 2026 chosen as the latest date available): https://www.bankofengland.co.uk/monetary-policy/inflation/inflation-calculator.

in 2024. Despite this being an even larger drop than in the UK (especially when the UK numbers are restricted to undergraduate enrolments), the situation in the US doesn't appear to be as urgent. Perhaps the main difference in the two systems is the way the US higher education market works, with universities having more freedom to increase tuition and fees. In 2012-13, the average annual tuition and fees were $8,070 and $24,523 for a public and private institution respectively. In 2024-25, that cost had risen to $10,153 and $37,937.[11] Even so, there are palpable pressures on US Philosophy departments, leading to recent closures or substantial cuts, including at Claremont Graduate (2018), Sonoma State (2025) and Toledo (2025). The losses are significant enough for the John D. and Catherine T. MacArthur Foundation to fund a project titled 'Investigating Cuts to Academic Philosophy' in 2025, with a report due imminently.[12]

The pressures felt by Philosophy departments are not unique but affect various other subjects. A report by the British Academy in 2025 highlights that "many parts of the UK are becoming [SHAPE, i.e. Social Sciences, Humanities and the Arts] subject cold spots".[13] In a worrying sign, they are even beginning to affect other subjects, e.g. STEM subjects, with some calling university degrees a waste of time and money. One recent survey, conducted by the Pew Research Center, found that 47% of Americans consider College worth it only if they don't have to take out loans to pursue it. Moreover, according always to the same survey, 49% consider a College degree as less important than 20 years ago (Fry, Braga and Parker 2024). To combat these trends, universities have been trying to adapt, but that sometimes means going for what they consider 'low hanging fruit' such as Humanities and Arts department closures or programme curtailments.

Another major pressure on universities is the emergence of much more capable AI technologies, particularly chatbots powered by large language models (LLMs). Chatbots are LLMs sheathed with an interface that allows humans to carry out conversations with them, as if we were talking to other fellow human beings. In the last four years, advanced chatbots have demonstrated uncanny abilities to fluently converse in natural language, deciphering human input and generating appropriate output that includes not only text, but also computer code, images and sound. Their use is spreading like wildfire. According to Ofcom's Adults' Media Use and Attitudes 2026 Report, in the UK, "[j]ust over half of adults (54%) say they use AI tools compared to three in ten (31%) last year" (5). Although the abilities of mainstream chatbots are generic and they generally suffer from 'hallucinations', people are increasingly relying on them to do everything under the moon, including work. As such, these tools further erode the perceived value of university degrees.[14] Some business leaders now think that all an employee needs to succeed, at least in certain roles, is prompting skills, i.e. extracting information from chabots through apt inputs (prompts). Others think that so-called 'soft' subjects are on the way out.[15]

A more direct threat to academics, including Philosophers, is AI's potential to take on teaching duties. The worst-case scenario sees academics being completely replaced by machines. Human tutoring, marking assignments and exams, as well as lecturing could all be a thing of the past in the future. John Warner writes that we should be as wary of the 'botification' of education, as we are of its 'adjunctification': "A PR release from the UCLA Newsroom about a comparative lit class that is using a 'UCLA-developed AI system' to substitute for labor… lays out the whole

[11] These figures are sourced from the Education Data Initiative that organises data from various existing sources such as the NCES: https://educationdata.org/average-cost-of-college-by-year.
[12] See https://www.futureofphilosophy.org/macarthur-announcement
[13] See https://www.thebritishacademy.ac.uk/policy-and-research/british-academy-shape-observatory/cold-spots-mapping-inequality-in-shape-provision/
[14] It doesn't help that a large segment of the population mistakenly thinks that university degrees are just for preparing people for the workplace.
[15] Interestingly, Karp has a BA in Philosophy, a JD and a PhD in Social Theory, all 'soft' disciplines.

deal. The course textbook has been generated from the professor's previous course materials. Students will interact with the AI-driven courseware. A professor and teaching assistants will remain, for now, but for how long?" (December 11, 2024, Inside Higher Ed).[16]

A similar story can be told with respect to research. AI and closely associated technologies are being used to create end-to-end automation. One landmark case, already a quarter of a century old, is the Robot Scientist, a machine that was able to formulate and revise hypotheses, design and conduct experiments, as well as interpret results in functional genomics (King et al. 2005). It arguably became the first machine to discover new knowledge independently of human help. More recently, the AI Scientist, "creates research ideas, writes code, runs experiments, plots and analyses data, writes the entire scientific manuscript, and performs its own peer review." (Lu et al. 2026). These are far from finished products, ready to replace academics in their entirety, but they do offer a glimpse of what's coming. Philosophical research is not immune to such developments.

To summarise, academic Philosophy is under growing pressure, at least in the UK and the US, with programmes and departments diminishing in size and even getting shut down. More generally, public perceptions of the value of university degrees, again at least in the US, seem to be increasingly negative. Even if universities survive this cocktail of volatility, academics still face a significant threat from AI. The progressively stronger teaching and research capabilities of AI-powered machines coupled with the cost savings they promise to deliver – being always on, unfailingly obedient, etc. – raise serious doubts about the long-term future of academics, Philosophers included. Having said this, it is the opportunities afforded by AI to Philosophy that this article is concerned about. We turn to this topic now.

4. Opportunities

In this Section we return to the methodological frameworks discussed in Section 2.2, asking how AI and computation can be exploited to boost research by carrying out tasks, helping us delve deeper into problems and solutions, scaling up the scope of our investigations, and opening up entirely new vistas of inquiry. An important distinction to keep in mind when reading this section is that between chatbots (or other user-friendly AI tools) and the models that underlie them. Both can be exploited in the aforesaid ways, but chatbots are easier to use while models offer more direct and fine-grained access to the encoded representations. The situation is analogous to asking a human a question versus accessing the representations in that human's brain via precise neuroimaging techniques. The former can be an easy and effective way to find out something about the human, but it won't always work as well as the latter, for several reasons, including dishonesty, limits to the power of introspection and ex post facto rationalisations.

As before, let us begin with dialogue. One obvious way to approach dialogues in the age of AI is by conversing with chatbots. This involves giving them clear instructions, a context and even a role. Indeed, context and role have emerged as important ways to engineer prompts that return more informative chatbot outputs (Sahoo et al. 2024).[17] Another way to exploit chatbots is by setting up multi-agent conversations. Agential AI is AI that is designed to have more autonomy, able to set and pursue goals, make decisions and take actions. In the case under consideration,

[16] Relatedly, there are well-known concerns about the threat AI poses to learning. When students pass off chatbot-generated content as their own, they are not only committing academic misconduct, namely plagiarism, but also eroding their own critical thinking and essay writing skills by failing to practice those skills. For a discussion of these and other downsides (as well as upsides), the reader may consult (Deep and Chen 2025).

[17] We conducted our own studies on prompt engineering, which we outline in the discussion of Experimental Philosophy below.

that means having the ability to initiate and carry out conversations, without following a strict one reply per prompt approach.[18] Needless to say, such agents can still be given instructions, a context, and even a role. Indeed, asking chatbots to role-play is particularly pertinent for dialogues, as that allows each chatbot to take on a different perspective in the conversation, making the dialogue more natural and interesting. Moreover, it allows us to co-ordinate debates between artificial counterparts of figures whose paths could not have crossed in actual history. For example, one can put Plato, Aristotle, Hume and Kant in one chatroom and let them debate about fundamental questions, such as the nature of reality. Moreover, one can run these debates thousands of times and study the resulting statistical properties to gain insight into the strengths and weaknesses of certain arguments and points of view.

Needless to say, the informativeness of these dialogues is modulated by the extent to which the LLMs that power them are any good. State-of-the-art LLMs are produced through Deep Learning (DL), i.e. learning algorithms that are applied to Deep Neural Nets (DNNs). DNNs are in some respects analogous to biological brains in that all the processing they do is accomplished via the sequential operation of connected neurons. Much like in our brains, the resulting models encode representations in a highly distributed fashion. LLMs, more specifically, are usually trained on vast amounts of linguistic data, allowing them to process and generate human-like language. Ultimately, they are statistical models that map linguistically convincing output to linguistically convincing input. Despite their usefulness and owing to the largely inscrutable nature of the built-up representations, LLMs suffer from various problems. These include lacking the relevant 'knowledge', not being able to provide cogent reasons for the issued claims, being prone to 'forget' the topic being discussed, being stubborn, exhibiting bias, and hallucinating (i.e. making up incorrect) answers.

Successive iterations of LLMs are getting better at dealing with these problems. Hallucination rates, for example, are on a downward trend (Bang et al. 2025). Moreover, we should be mindful that humans also suffer from comparable problems, e.g. being overconfident, making stuff up when we don't really know, and not being able to give good reasons for why one ought to adopt our point of view. The bar should thus not be set too high for LLMs, at least when deciding whether to use them for non-critical research. There are other ways to mitigate some of the aforesaid problems. For example, instead of generic LLMs, one can opt for specialised LLMs that have been either fully trained or fine-tuned. Fully trained models are models that are trained from scratch. The upside is potentially increased performance, i.e. improved mimicry of human knowledge, skills, and even writing style vis-à-vis the domain of interest.[19] The downside is that training requires massive specialised datasets and massive compute power. Fine-tuned models are not trained from scratch. Instead, a pre-trained (often generic) model is adjusted by providing additional training on a specialised but smaller dataset. The upside is, once again, increased performance (vs. generic models) vis-à-vis the domain of interest, but now the training no longer requires massive specialised datasets or massive compute power. The downside is that any performance increase is likely to be modest compared to fully trained specialised models.[20]

---

[18] For a recent low-level example of multi-agent interactions, see David Freeborn's 'How AI Agents Handle a Traitor in Their Midst?' experiment reported in a post on our blog: The Computational Philosopher blog.

[19] There is some evidence (e.g. Nori et al. 2023) to suggest that, contrary to common sense expectations, generic LLMs outperform specialised LLMs. We counsel caution when reading such studies. A fair comparison would have to ensure that all other things are equal, e.g. using the same DNN architecture in both types of LLMs. Moreover, it would have to ensure that the specialised LLMs are not overly narrow, as otherwise critical knowledge, skills and styles for the specialised tasks would not be learned, thereby defeating the purpose of producing such models.

[20] For a striking example of how machine learning and natural language processing can be used to detect styles of writing (stylometry) which can influence philosophical and historiographical hypotheses, one

In the case at hand, one could fully train a Philosophy Model from scratch using the entire PhilPapers database (c. 140k articles), plus the Stanford Encyclopedia of Philosophy (c. 2k entries) and the Internet Encyclopedia of Philosophy (c. 1k entries).[21] We estimate this would be sufficient to train a mid-sized (e.g. 70B parameter) model in 2026. Fine-tuning models is less data- and compute- hungry. One can take a pre-trained model from HuggingFace and fine-tune it with a custom dataset, including the aforesaid encyclopaedias, using specialised tools like Unsloth or Ollama.[22] The pre-trained model could be as large as the largest models out there, currently a few trillion parameters, or as small as a few hundred million parameters. One example of a fine-tuned model in Philosophy is Digi-Dan, a model trained on Dennett's writings using GPT3 as a basis. Digi-Dan had some success in fooling experts (Schwitzgebel, Schwitzgebel, and Strasser 2023).[23] These days one can even fine-tune small to mid-sized models at home, provided they have the right hardware.[24] Needless to say, the potential usefulness of fully trained or fine-tuned specialised models applies to what we say about other methodological frameworks.

Let us consider next the methodological framework of thought experiments. One obvious way to approach thought experiments in the age of AI is to instruct chatbots to produce them. To increase the chances of getting something meaningful, it is important to provide clear instructions, a context and perhaps even some examples of existing thought experiments as a template. Here's a simple prompt we tried out: "Using the Chinese Room Argument as a template, construct a thought experiment that argues for the superiority of structural realism over standard scientific realism". To give a little context, structural realism, roughly speaking, holds that we can only know the structure of the world, i.e. the relations between things but not the intrinsic properties of those things. Standard scientific realism rejects that limitation.[25] After several additional prompts to help adjust the output, as well as some admittedly appreciable manual corrections, here's the text we ended up with: 'A human engineer is kidnapped by aliens and confined to a room on another planet. They are tasked with creating a map of that world and given all sorts of raw materials and tools. Their only recourse is to build measurement-collecting probes, small enough to fit through a hole in the room. Their measurements are, by necessity, made in relation to each other and to the room they use as their base. Through these measurements the engineer slowly builds up a map of this world. The map is perfectly accurate, yet the engineer's only knowledge cannot but be purely relational, i.e. structural. Much like the engineer, scientists never have access to absolute points of reference like intrinsic properties. All their measurements are purely relational. Their knowledge must thus also be purely structural'. In Searle's Chinese Room thought experiment, the punchline is that symbol manipulation is insufficient for understanding beyond syntax, i.e. meaning. Here the punchline is that measurement is insufficient for knowledge beyond structure, i.e. intrinsic properties. Whether the chatbot-generated thought experiment is at all compelling is up to the reader to decide. All that matters for our purposes is that such experiments can in principle be generated via chatbots.

---

may consult Schmidt-Petri, Schefczyk & Osburg (2022). This presents a computationally-driven stylometric analysis of J.S. Mill's *On Liberty*, helpfully suggesting which parts are likely to have been written by Harriet Taylor, his wife. That his wife contributed to this and other J. S. Mill works is rather well known, but which parts were hers is still debatable.

[21] Some of my colleagues have been exploiting NLP to analyse the use of language in AI Ethics papers – see Cline et al. (2025). They ultimately want to analyse huge databases like PhilPapers.

[22] HuggingFace is an online library of shared LLMs, datasets and tools.

[23] See also the Peter Singer AI chatbot: https://www.petersinger.ai/

[24] We fine-tuned a 4B parameter LLM on our own writings at home, using a 16GB Nvidia RTX graphics card. The fine-tuning didn't take long. Getting the data ready, e.g. cleaning and structuring it, was much more time-consuming.

[25] Alas, there is no space here to discuss structural realism in detail. The interested reader may consult Frigg and Votsis (2011).

LLMs specialising in thought experiment construction, either through full-training or fine-tuning, would naturally be expected to be more compelling. Training from scratch is markedly more difficult in this case because of the scarcity of thought experiments. Still, there are various ways around this problem, though these are neither easy nor straightforward. Consider first an analogous case. Raven's progressive matrices are employed in tests of fluid intelligence (Burke 1958). One is asked to inspect a row of shapes and then predict the missing shape in the next few rows. Training a model on a DNN architecture to accurately make such predictions is notoriously difficult because there just aren't enough such human-crafted examples of progressive matrices for the model to learn from. To solve this problem, data scientists have been devising algorithms that generate novel examples of matrices that can then be employed for training (Mitchell 2021). The question that immediately arises is whether the synthetically-generated matrices are legitimate, and, moreover, whether they are representative of the space of matrices as a whole. There is a catch-22 problem here: an algorithm that generates perfectly legitimate and representative matrices would presumably already contain the knowledge that those DNN-produced models are aiming to acquire. But that knowledge is precisely what we are missing. Similarly with our case, having an algorithm that generates perfectly legitimate and representative thought experiments would defeat the purpose of having to train an LLM from scratch to do the same. Still, it's not all doom and gloom. An algorithm may create enough examples to provide a useful crutch in such training. Needless to say, the potential usefulness of synthetically-generated data applies to what we say about other methodological frameworks.

Next on the list of methodological frameworks is conceptual analysis. Vincent Müller, in a recent talk, highlights the prospects of utilising AI, specifically chatbots, to test philosophical hypotheses about concepts. The upshot of such an approach, according always to him, is the avoidance of anthropocentric bias. We agree that AI can help test philosophical hypotheses but would like to point out two ways in which our thoughts and Müller's diverge. First, LLM-powered chatbots are very much anthropocentric. That's because their models are trained on masses of human linguistic data. That also helps explain why, when left alone (without training or other interventions), they not only reproduce human biases, but also sometimes amplify them – see Gallegos et al. (2024). Second, conceptual analysis is not just about justification, but also about discovery. In what follows, to show off both the justification and discovery abilities of chatbots, we conducted a proof-of-concept investigation.

Let us start with justification. We prompted Claude (version: Sonnet 4.6) to "[l]ist commonly held intuitions about knowledge that would be useful to account for by a definition of knowledge". It provided a list of about 30 such intuitions, including (rewritten here for brevity): knowledge must be factive, it requires belief, it must be more than belief, it must not be mere luck, it must be possessable by animals, it must exclude serious doubt, it must have justification, the justification must not be accidental, the strength of the justification matters, and justification and knowledge must be linked in the right way. In the same session, we instructed the chatbot to test philosophical hypotheses about the concept of knowledge, using the following prompt "Taking into consideration the commonly held intuitions about knowledge you listed above, which, if any, definition of knowledge comes closest to meeting them?". The reply involved a comparison of various theories, including justified true belief, reliabilist, sensitivity, safety, virtue epistemology, contextualist, and knowledge-first accounts. Claude qualified its overall assessment by asserting that "no theory fully satisfies all the intuitions". Nonetheless, it singled out 'safety-based virtue epistemology' as the most promising, citing its ability to do justice to intuitions of reliability, animal knowledge, and modal luck, among others. The rationale provided was not perfect, but it made enough sense to potentially be of use in a systematic philosophical investigation conducted by humans.[26]

[26] By the way, this was despite the fact that we put Claude in a 'low effort' setting.

Consider next discovery. We prompted Claude to "[t]ry to take into account all thought experiments relating to theories of reference and then postulate an original definition of reference". It first identified what it branded as 'The Inherited Landscape', listing and briefly discussing Frege's puzzle, Russell's theory of descriptions, Kripke's modal argument, Kripke's epistemic argument, Putnam's twin earth thought experiment, Putnam's division of linguistic labour, causal-historical chains, Donnellan's referential/attributive distinction, Evan's causal theory and its problems, Searle's cluster theory, Burge's social externalism, Wittgenstein's rule-following, Kripkenstein, Evans' descriptive names, 2D semantics and mental files theory. It then posited what could best be described as an 'equilibrium theory' of reference: "Reference is the stabilization of a tracking relation between a linguistic expression and a worldly entity, achieved through the dynamic tension between causal-historical origin, communal normative governance, and the ongoing indexical anchoring of use to world". As made clear by the quoted text, this theory incorporates concerns about the expression's initial baptism, continued use and indexical context, seeking to identify reference via some sort of stabilisation process driven by those concerns. Once again, the result is far from perfect, but may potentially be of use in a systematic philosophical investigation conducted by humans.

Deploying chatbots and other AI technologies to assist the justification and discovery dimensions of conceptual analysis is highly promising, especially as these technologies mature and more specialised models are released. The fact that these models are trained on linguistic data is highly pertinent to the whole premise of conceptual analysis, namely that we test and devise hypotheses about concepts based on how language is being used, or, how it should be used. Additional constraints are often advisable. One such constraint has to do with the fact that no concept is born in isolation. We are always dealing with clusters of co-dependent concepts, where each concept's meaning affects several others, often to different degrees.[27] We thus need to opt for definitions or characterisations of concepts that maximise coherence within, and even across, clusters of concepts. That's something that humans already do but at a small scale, e.g. a handful of concepts at a time. The kind of scale that's required to co-define/co-characterise dozens, hundreds or even thousands of concepts simultaneously cannot but be achieved with machines, able to check how each modification of a concept affects numerous others. Automating this process would be a game changer, with implications far beyond Philosophy.

Moving on to meaning-as-use, this methodological framework can benefit from AI and related technologies for some of the same reasons mentioned in our discussion of conceptual analysis. The aim may no longer be to devise proper definitions, opting instead for fuzzier concepts centred around prototypes, but the key idea is still the same. Indeed, we cannot think of a better fit between a methodological framework and the way language models are trained as in both cases what is paramount is actual language use. To demonstrate what's possible, let us first review two central ideas in natural language processing (NLP), namely word embeddings and semantic similarity metrics.

Put simply, word embeddings are numerical representations of the meaning of words. A little more precisely, word meaning is represented by vectors, i.e. sequences of numbers. For example, the meaning of the word 'cow' may be represented by the vector [1.3, -0.2, 5.6], which contains three numbers. What are these numbers? Each represents a different dimension – think property

[27] This is, of course, known as the thesis of semantic holism. Holism, including semantic holism, comes in various shapes and forms. We assume a moderate version of semantic holism, whereby not all concepts are directly connected to all others, and, moreover, not all connections have the same strength. It's also noteworthy that the thesis of semantic holism is associated with the thesis of distributional semantics. See below for a discussion.

– of word meaning. The dimensions may not be human scrutable, but a machine's way of latching on to the nuances of meaning. Why *three*? That number is just for illustration purposes. Current wisdom holds that to make embeddings useful about *three hundred* dimensions are needed. Where do we get these dimensions? We get them from word embedding models, e.g. Word2Vec and Glove, that exploit the thesis of distributional semantics, expressed neatly and concisely by the maxim: 'You shall know a word by the company it keeps' – see Harris (1954) and Firth (1957). These models are trained on linguistic data, e.g. Wikipedia articles, and track how frequently words co-occur, e.g. the words 'dog' and 'cat' often co-occur with the words 'pet' and 'fur'. More sophisticated embedding models, e.g. BERT, are context-sensitive, assigning different vectors to the same word based on context. Finally, state of the art LLMs, e.g. GPT and Claude, take this a step further, assigning probabilities to sequences of tokens and using those probabilities to generate text. To summarise, word embeddings allow us to locate the meaning of a word in unified multi-dimensional spaces.

This brings us to the next central idea in NLP, namely semantic similarity metrics. Our confidence in a model's word embeddings increases when similar words appear near each other in the aforesaid multi-dimensional space. Thus, we expect 'dog' and 'cat' to be nearer to each other than, say, 'roses' and 'quarks'. Moreover, our confidence increases when the geometry of that space reflects semantic relations we deem intuitive. A well-worn example involves the words 'king', 'queen', 'man' and 'woman'. If we subtract the vector of the word 'man' from the vector of the word 'king' and then add the vector of the word 'woman', we should ideally get a vector in that space that's close (indeed, closest) to the vector of the word 'queen'. To determine how close a word is to another word in the multi-dimensional space requires a similarity metric. Some well-known distance metrics are employed to play this role. They include Euclidean distance, dot product, Manhattan distance, and cosine distance. Besides word similarity, there are also sentence and document similarity metrics. As their names suggest, these allow for semantic comparisons between sentences and even whole documents.

Let us return to the meaning-as-use methodological framework for some potential applications that exploit the above NLP ideas. One obvious application involves tracking changes in meaning through their usage in relevant corpora. For example, one could look at whether, and if so how, the words 'good' and 'evil' may have changed over time, testing Nietzsche's claims in the process, by considering their closest neighbours in the corresponding historically contextualised multi-dimensional spaces. Another application involves pinpointing prototypical concept tokens by finding their centroids, i.e. the most central tokens in clusters. For example, one could investigate which, if any, conception of 'free will' deserves to be crowned as prototypical in Metaphysical discussions, do the same in Moral discussions and compare the two along the way. This may also help clarify whether some authors talk past each other. Yet another application involves determining concept boundaries by checking where token similarity to a prototype drops off dramatically.[28] For example, the concept 'heap' may be determined to be inapplicable in cases where less than three objects are involved, precisely because the distance to its vector from the vectors of concepts zero, one and two is substantially longer than to any other vector of a natural number, according to some acceptable similarity metric(s).

To give a more detailed illustration of what can be achieved with NLP methods, consider an ongoing project that tracks meaning shifts in Darwin's language – see McGillivray, Lyons, Smith,

[28] The term 'token' is here used in a different way to the way it is used in NLP. In prototype theory, tokens are just concepts that fall under other concepts we call types. For example, 'robin' is a token as it falls under the type 'bird'. In NLP, tokens may be words, parts of words, or even characters. Moreover, in context-sensitive models, e.g. BERT, different instances of the same word count as distinct tokens. For example, the sentence 'Robins fly but robins are small' contain two tokens of the word 'robin'.

Smithies & Votsis (under review). Word-embedding models, namely Word2Vec and BERT, are trained on Darwin's outgoing correspondence. The ensuing NLP analysis, which utilises the cosine similarity metric, suggests that there are both gradual and abrupt shifts in Darwin's language, with some words, e.g. 'Linum' (a genus of flowers), undergoing significant changes over time. One indicator of abrupt shifts is the change observed in a word's top neighbouring words. The analysis also suggests that semantic shifts are more pronounced when Darwin is addressing a scientific (vs. non-scientific) audience, perhaps as a reflection of the increased semantic negotiation taking place between scientists in a developing field. Finally, the analysis suggests sources of influence in Darwin's work. A fine-tuned version of Mistral, an LLM model, was tasked to identify conceptual continuities between Darwin's outgoing correspondence and his publications, looking at entire documents, not just words. Of all of Darwin's correspondents, Joseph Hooker seems to have had the greatest impact, as indicated by the closest semantic similarities between Darwin's letters to Hooker (compared to any other correspondents) and Darwin's publications. It is worth emphasising that the above suggestions are tentative, pending further examination.

Next up let us consider the methodological framework of explication. What we need from the computational and AI toolkit are methods that help us satisfy Carnap's four desiderata: similarity to the original concept, exactness, simplicity and fruitfulness. The first one is straightforward. As we saw earlier, semantic similarity metrics can be employed to give us some clues about the closeness in meaning between an explicated and the original concept. The second desideratum is perhaps a little more involved. To determine how exact a concept is, we may need to employ measures/tools that track the internal consistency, or, more broadly, the internal coherence of a concept such as description logic satisfiability checking (Baader 2009), targeting the presence of contradictions (ideally none), and within-class variance of concept token similarity, targeting meaning disunity between concept tokens (the lower the better). The third desideratum can also be computationally addressed. Syntactic simplicity, for example, can be measured by minimum description length methods (Rissanen 1978), which effectively look at the compressibility of the representation. The final desideratum, fruitfulness, may also be measured in numerous ways. We explore one such way here, namely predictive utility. Definitions that increase the predictive utility of hypotheses employing them can be judged to be more fruitful. To test this, one can resort to an automated theorem prover, or symbolic reasoner more generally, to compare the rates of true and non-trivial consequences/predictions of rival hypotheses whose only difference are the divergent definitions of a key concept. Suppose we define concept X as 'neutrally charged leptons'. The hypothesis 'In charged-current interactions with X, electrons are produced' will only make correct predictions one third of the time. If, instead, we define X as 'neutrally charged leptons with a mass less than 0.8 eV/c^2', the resulting hypothesis makes correct predictions all the time. The reason is that neutrally charged leptons with a mass less than 0.8 eV/c^2 correspond to electron neutrinos, which unlike muon or tau neutrinos (which are also neutrally charged leptons but have different masses), are the only ones producing electrons in charged-current interactions.[29]

The second to last methodological framework to be explored here is reflective equilibrium. Besides asking chatbots to conjure up intuitions and principles, one can automate the process of iteratively modifying each to achieve equilibrium. In fact, the search for an equilibrium between intuitions and principles can handily be treated as an optimisation problem. The optimal state of such an equilibrium is when we find a version of a principle that stands in a maximal explanatory coherence relation to some version of the intuitions. An interesting formal model of reflective equilibrium, partly computationally implemented through simulations, can be found in Beisbart, Betz and Brun (2021). On this model, to reach an equilibrium involves the optimisation of an

[29] This example assumes, for simplicity, that there is an equal ratio of the three types of neutrinos when predictions are made with the hypothesis where X is defined in terms of 'neutrally charged leptons'.

achievement function that combines the criteria of accounting (the extent to which principles agree with intuitions via entailment), systematicity (the extent to which principles are simple) and faithfulness (the extent to which the modified intuitions are continuous with the initial intuitions).

No matter what formal models we use, the process can be boosted with computational and AI methods. What one needs are metrics for the notion of explanatory coherence and algorithms that alternately adjust intuitions and principles in a principled way until a stable state is reached. To give an illustration of what is possible, consider how symbolic AI methods may help. Explanatory coherence is often rendered, at least partly, as a non-trivial inferential relationship between beliefs. In the case at issue, such relationships would hold between principles and intuitions, both expressed as statements. Symbolic AI excels at this sort thing. We here focus on a symbolic AI approach that can also learn from data, known as Inductive Logic Programming (ILP). This approach exploits both inductive as well as deductive logic and can be used to search the space of possible principles. To find a principle P, based on background knowledge B, positive intuitions I^+ and negative intuitions I^-, ILP can utilise algorithms whose aim is to find the least general generalisation P such that $B \wedge P \models I^+$ and $B \wedge P \not\models I^-$. ILP can also help with the principled adjustment of P and I. We just need another layer of complexity to achieve this. For example, one can shift intuitions that are deemed undesirable, perhaps because they lack independent motivation, from I^+ to I^-, but also replace the intuitions shifted from I^+ with modified versions thereof by expanding or contracting the extension of their concepts – see the discussion above. Iterating this process, allows us to formulate different candidate Ps until one is reached that can satisfy $B \wedge P \models I^+$ and $B \wedge P \not\models I^-$ and any other constraint, e.g. fruitfulness, we throw at it.

The last methodological framework to discuss here is X-Phi. Read strictly, this approach involves experiments that test the extent to which intuitions are shared in various populations. A simple way to infuse this framework with computational methods is by using Agentic AI chatbots to tease out intuitions through natural conversation with humans. A closely related approach utilises Agentic AI chatbots to generate novel thought experiments vis-à-vis some concept or category of interest that are then fed to humans via an online platform. The human responses to the chatbot questions can then be statistically analysed to determine to what extent the underlying intuitions lean one way or another. Yet another related approach involves two types of Agentic AI chatbots. The first type converses with the second by posing questions on specific topics, with or without generating thought experiments. The second type replaces the human participants by thousands of independent Agentic AIs that generate responses to those questions. On the assumption that the second type of Agentic AI is a good stand in for humans, we may be able to learn something useful about the dis-/uniformity of intuitions and the theories they presumably underwrite.

That assumption can, of course, be questioned. Indeed, there is some (admittedly very preliminary) evidence to suggest that current LLMs produce content that might be generally more uniform than the content produced by humans. In a recent project (Votsis, Freeborn, Ball and Norelli draft), we analysed human-to-human interactions and human-to-chatbot interactions to find out whether the way a question is written affects answer quality. Our data consisted of human questions and answers about simple Python coding problems taken from the Stackoverflow forum as well as chatbot answers produced when the aforesaid human questions were fed to them. Our analyses revealed correlations between various textual metrics.[30] Crucially for the current issue being discussed, we found that chatbot answers were more uniform in quality than those of humans. Moreover, it is worth keeping in mind that a chatbot's distribution

[30] For example, in the case of human-to-human interactions, we found that polite, easy-to-read questions tended to produce more readable answers. In the human-to-chatbot interactions, more grammatical/readable questions produced more grammatical/readable answers. We also found that chatbots were more sensitive to the way a question was written.

of responses need not correspond to any particular human population.[31] As such, attempts at implementing the two-type Agentic AI approach to X-Phi may woefully miss the mark. We are not suggesting here that this challenge is intractable. After all, one solution to it is to fully train or fine-tune several LLMs, each of whose responses reflect a distinct human population.

A completely unrelated approach to doing X-Phi with AI employs NLP but not chatbots. This approach involves scouring the internet, e.g. online fora, social media, etc., for instances of the concepts or categories of interest, extracting the surrounding text and using a trained or fine-tuned model to determine the most likely intuition being endorsed. For example, a comment like 'Thinking is not the same as knowing' may be taken to endorse the intuition that 'knowledge is more than belief (where believing = thinking)'. Provided enough data and their provenance can be gathered this way, we may be able to build a good map of which human populations endorse which intuitions, and, by extension, which theories of concepts are best supported, at least for the specific populations, if not for humanity as a whole.

Before we bring this section to a close, it is worth qualifying what is being said here. First, none of the above computational / AI implementations of the discussed methodological frameworks are without problems. Existing semantic similarity metrics, for example, are not as nuanced, at least not across the board, as human judgment, particularly not expert human judgment. Still, there is enough capacity present in those methods for useful work to be carried out and enough validation has been performed on them to inspire guarded confidence. Moreover, these methods are continually improving – think, for example, of the move from static to contextual word embeddings. Second, many of the examples of computational / AI implementations are not tied to the methodological frameworks they were offered with, but, *mutatis mutandis*, apply more generally. For example, the use of symbolic AI to determine whether a concept is free of internal contradictions in explication, can find more widespread application by checking the validity or persuasiveness of reasoning in dialogues and thought experiments. Finally, this article does not fetishise the specific computational and AI technologies, tools and methods discussed. Most of these are understood to be transient, to be replaced before long by more powerful versions and even entirely novel developments in computation and AI. A case in point is LLMs. These are only 'intelligent' by piggybacking on text. That is, they learn to statistically mimic the linguistic symptoms of human intelligence, which is itself grounded in worldly interactions. As Yann Le Cun points out in a recent interview in the Financial Times, LLM days are numbered (Heikkilä 2026).

5. A Few Thoughts on the Future

Clearly, the aforesaid AI threats and opportunities are closely related to each other. The very same computational and AI technologies that can turbo-charge the work of Philosophers can also help proliferate errors and bias as well as eventually make human contributions redundant. An extreme scenario sees not only humans being replaced by machines, but also ultimately the subject of Philosophy deemed supererogatory and dissolved as an area of active research. Whether or not the overall effect of computational / AI technologies to Philosophy is positive, depends on various implementation details. In this respect, they are not unlike most other technologies, though they are obviously more on par with nuclear technology in the existential threat they pose. At the moment, there is no question of wholesale replacement of all Philosophers by machines. Rather, what's at stake is the replacement of certain tasks, call it 'partial replacement', Philosophers carry out, which may lead to some reductions in the workforce (e.g. a result of needing less human time to perform related teaching, research or admin duties). In what follows, we offer some tentative remarks about which core and primary areas of Philosophy research are likely to be impacted the most and why.

[31] I am grateful to the one of the referees for reminding me of this point.

We expect research in some areas to see fairly rapid and widespread adoption of computational or AI technologies. Let us begin with the core areas. It is our contention that most (if not all) of these will soon be significantly impacted by computational or AI technologies. Indeed, we expect a computational/AI turn akin to that which occurred with the advent of naturalised Philosophy.[32] In effect, most research in those areas will continue as before, but a non-negligible and steadily growing slice will be computational/AI-driven. We are already seeing glimpses of this with the use of computer simulations in Epistemology (Ball et al. 2024; Zollman 2015), automated theorem provers in Metaphysics (Fitelson and Zalta 2007; Benzmüller and Woltzenlogel Paleo 2017), and NLP in Ethics (Bystranowski, Dranseika and Żuradzki 2022; Zangari et al. 2025). The opportunities to make progress on such questions as how beliefs ought to be justified, what categories and foundational and whether there are moral values shared across cultures will prove irresistible.

Moving on to the primary areas, it is our contention that research in most (if not all) of these will also soon be significantly impacted by computational or AI technologies. As primary areas are more specialised than core areas, it is more likely that some primary areas will in fact end up being heavily reliant on those technologies. Candidates include the Philosophies of Language, Art and Science. Take the Philosophy of Language first. Several of the abovementioned methodological frameworks, e.g. conceptual analysis, meaning as use, and explication, that, as we just saw, are ripe for a computational or AI transformation are also intimately tied to, and indeed emerged from considerations in, the Philosophy of Language. Moreover, the corresponding scientific discipline, i.e. Linguistics, is fast becoming dependent on computational and AI tools (Alduais, Yassin and Allegretta 2025). Indeed, all scientific disciplines are inexorably moving towards computational and AI methods (Sajadieh et al. 2026). That in itself is reason enough to bet on the Philosophy of Science's research methodology being transformed by those technologies. Add to it the fact that several debates in this area have been making practical recommendations for years, e.g. confirmation theory (Tentori, Crupi, Bonini and Osherson 2007) and scientific discovery (Votsis 2024), the opportunity to finally translate these recommendations into something tangible with the help of computational and AI technologies will increasingly become hard to pass by. As for the Philosophy of Art, the torrent of AI-generated art flooding the market is plain for all to see, for beyond its effects on the job market, it raises important philosophical questions about creativity and value (Helliwell 2026). Carrying out research experiments with the assistance of computational and AI technologies to throw light on those questions will soon become a familiar sight.

To be clear, we are not claiming that the aforesaid methodological frameworks will disappear. Rather, we are claiming that finding computational or AI implementations of these frameworks will become more likely than not. Moreover, we are not claiming that only positives will ensue from this transformation of research. The most likely scenario is that we will see a mixture of positives and negatives. At first, most work carried out with such methods will be clumsy and unsophisticated. Take, for example, semantic analyses performed via NLP. Some word embedding models have been known to place the words 'expensive' and 'inexpensive' near each other in the vector space because they frequently appear in similar contexts (cf. Mrkšić et al. 2016). This is clearly wrong as the two words are antonyms. Triangulating meaning is a highly challenging problem. First off, one needs to latch on to the relevant statistical usage patterns. But even that may not be enough. As various philosophers have insisted, meaning is not exhausted by such patterns since there are also real-world connections to the way we use language. The word 'sunset', for example, is not just captured by its co-occurrence with other words, e.g.

[32] Note that the latter happened over several decades. For example, the naturalisation of Epistemology took off in the 1960s, whilst the naturalisation of Metaphysics had to wait another four decades. We expect much shorter timelines in the case of the computationalisation/AI-fication of Philosophy.

'beach', but also arguably by its connections to events such as the Sun setting over the horizon and the experiences those events induce in us.[33]

These, and other challenges across the computational/AI body of methods, are not insuperable. For example, the shift to richer embedding models, capable of more contextual nuance, as well as to AI agents with sensori-motor abilities, capable of real-world connections, is already going some way towards improving these methods. Moreover, lest we forget, human beings are themselves confronted by many of these or similar challenges. One needs to simply have a look at the vast amounts of literary and scientific slop we produce, our endless re-dressing of ideas with different clothes and passing them off as something completely novel, and our propensity to focus on the banal, as glaring evidence of how challenged we truly are. And while our progress in addressing these challenges has been slow, our greatest success has been in creating machines that may one day be able to completely overcome their own challenges and ours. Indeed, from where we sit, the march of the machines towards meeting those challenges seems unstoppable. We are optimistic that in the medium term, the positives brought about by utilising machines to augment our work will far outweigh the negatives. Some such work will even become seminal in the history of Philosophy.

It is finally worth speculating, what will happen in a longer-term future where machines are able to conduct Philosophical research at vastly increased speeds, scales, accuracy, precision, and neutrality, compared to humans. In such a future, would it not be irrational to allow machines or machine-augmented humans to conduct end-to-end Philosophical research. Would there even be a point to continue to involve un-augmented humans (if any such remain)? Clearly, not if results are what matter most. However, if Philosophy is pursued in an Emersonian 'it's not the destination, it's the journey' that counts, then such humans may yet find themselves philosophising for years to come. The psychological benefits of discovering things for oneself may even be recommended to un-augmented humans by future machine benefactors (or overlords), as a way of combating the human condition (or to rationally pacify us).

6. Conclusion

Academic Philosophy is at a crossroads facing unprecedent threats, but also opportunities, especially those emanating from AI. In this paper, we directed our attention to the opportunities, especially those relating to the subject's methodological frameworks. We explored seven such frameworks (dialogue, thought experiments, conceptual analysis, meaning as use, explication, reflective equilibrium and experimental philosophy) and gave examples of how they can be transformed by computation and AI. The ability to offload menial tasks, precisify problems and solutions, increase the speed, depth and scale of research, discover patterns that are difficult (perhaps even impossible) for humans to see, as well as open entirely new vistas of inquiry are too good a set of opportunities to turn down. Our projection is that in the short-term, the computational/AI implementation of methodological frameworks in Philosophy will lead to clumsy and unsophisticated work, but that will change over time, say in the next twenty years. The infusion of Philosophical methodology with AI and computation will not only become dominant, at least across several primary areas, but will also lead to progress and seminal works.

Perhaps the most significant opportunity afforded by computation and AI to academic Philosophy is a stay of execution, perhaps a permanent one. As AI and computational abilities as well as public interest grow, so will Computational Philosophy and Philosophy of AI programmes. Indeed, unless the funding model of universities changes radically, attracting students into those programmes will be the safest bet for the survival of the subject. This will also have downstream effects. Future Philosophers will need to possess relevant AI expertise, if they are to be hired by

[33] Putnam's well-known dictum 'meaning… ain't [just] in the head' is particularly apt here.

departments to teach those students, but also to conduct research in those areas. Without such expertise, they will likely have more trouble getting their research funded, impacting on their career prospects. As for the Philosophers of today, we have a duty to prepare students for that future, circumnavigating various threats whilst steering them towards undeniable opportunities.

References:

Alduais, A., Yassin, A. A., & Allegretta, S. (2025). Computational linguistics: a scientometric review. *Quality & Quantity*, *59*(5), 4097-4136.

Ball, B., Koliousis, A., Mohanan, A., & Peacey, M. (2024). Misinformation and higher-order evidence. *Humanities and Social Sciences Communications*, *11*(1), 1294.

Bang, Y., Ji, Z., Schelten, A., Hartshorn, A., Fowler, T., Zhang, C., ... & Fung, P. (2025, July). Hallulens: Llm hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 24128-24156).

Beisbart, C., Betz, G., & Brun, G. (2021). Making reflective Equilibrium Precise: a formal model. *Ergo* 8 (15): 441–472.

Benzmüller, C., & Woltzenlogel Paleo, B. (2017). Experiments in Computational Metaphysics: Gödel's Proof of God's Existence. *Savijnanam: scientific exploration for a spiritual paradigm. Journal of the Bhaktivedanta Institute*, *9*.

British Academy. (2025, September 10). Cold spots: Mapping inequality in SHAPE provision in UK higher education. https://www.thebritishacademy.ac.uk/policy-and-research/british-academy-shape-observatory/cold-spots-mapping-inequality-in-shape-provision/

Burke, H. R. (1958). Raven's Progressive Matrices: A review and critical evaluation. *The Journal of genetic psychology*, *93*(2), 199-228.

Bystranowski, P., Dranseika, V., & Żuradzki, T. (2022). Half a century of bioethics and philosophy of medicine: A topic-modeling study. *Bioethics*, *36*(9), 902-925.

Carnap, R. (1950). *Logical Foundations of Probability*, first edition, Chicago: University of Chicago Press.

Cartwright, N. (2004). Causation: One word, many things. *Philosophy of Science*, 71(5), 805-819.

Cline, A., Ball, B., Freeborn, D. P. W., Helliwell, A. C. and Loi-Heng, K. (2025), 'Computational Analysis for Philosophical Education: A Case Study in AI Ethics', *Edukacja Filozoficzna*, 79, pp. 47–74. doi: 10.14394/edufil.2025.0003.

Cooper, J. M. (ed.) (1997). *Plato: Complete Works*. Indianapolis: Hackett.

Deep, P. D., & Chen, Y. (2025). The role of AI in academic writing: Impacts on writing skills, critical thinking, and integrity in higher education. *Societies*, *15*(9), 247.

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 10-32.

Fitelson, B., & Zalta, E. N. (2007). Steps toward a computational metaphysics. *Journal of Philosophical Logic*, *36*(2), 227-247.

Freeborn, D. (2026) 'How AI Agents Handle a Traitor in Their Midst?', Blog Post, *The Computational Philosopher*, https://computationalphilosopher.substack.com/p/how-do-ai-agents-handle-a-traitor.

Frigg, R., & Votsis, I. (2011). Everything you always wanted to know about structural realism but were afraid to ask. *European journal for philosophy of science*, *1*(2), 227-276.

Fry, R., Braga, D. and Parker, K. (2024) 'Is College Worth it?', *Pew Research Center*.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational linguistics*, 50(3), 1097-1179.

Goodman, N. (1965). *Fact, Fiction, and Forecast*, second edition, Indianapolis, IN: Bobbs-Merrill.

Guyer, P., and Wood, A. (eds.) (1998). *Critique of Pure Reason*. Cambridge: Cambridge University Press.

Harris, Z. S. (1954). Distributional structure. *Word*, 10 (2-3), 146-162.

Heikkilä, M. (2026). Computer scientist Yann LeCun: 'Intelligence really is about learning'. Financial Times, January 2026.

Helliwell, A. C. (2026). Can AI Create with Flair? Artistic Creativity and Artificial Intelligence. *British Journal of Aesthetics*, ayaf057.

Jackson, F. (1998). *From metaphysics to ethics: A defence of conceptual analysis*. Oxford University Press.

King, R. D., Young, M., Clare, A. J., Whelan, K. E., & Rowland, J. (2005). The robot scientist project. In *International Conference on Discovery Science* (pp. 16-25). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kornblith, H. (2002). Knowledge and Its Place in Nature. New York: Oxford University Press.

Lu, C., Lu, C., Lange, R. T., Yamada, Y., Hu, S., Foerster, J., ... & Clune, J. (2026). Towards end-to-end automation of AI research. *Nature*, *651*(8107), 914-919.

McGillivray, B., Lyons, S., Smith, E.L., Smithies, J. & Votsis, I. (under review). Darwin's Language Evolution: decomposing computational methods in the age of AI.

Mitchell, M. (2021). Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, *1505*(1), 79-101.

Mrkšić, N., Séaghdha, D. O., Thomson, B., Gašić, M., Barahona, L. M. R., Su, P. H., ... & Young, S. (2016, June). Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 142-148).

Munis, J. (2026). 'Palantir CEO says AI 'will destroy' humanities jobs, but there will be 'more than enough jobs' for people with vocational training', Fortune, April 11 2026, https://fortune.com/article/palantir-ceo-alex-karp-ai-humanities-jobs-vocational-training/.

Newell, A., & Simon, H. (1956). The logic theory machine--A complex information processing system. IRE Transactions on information theory, 2(3), 61-79.

Nori, H., Lee, Y. T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., ... & Horvitz, E. (2023). Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

Ofcom (2026). Adults' Media Use and Attitudes Report.

Perry, J. (1978). *A dialogue on personal identity and immortality*. Hackett Publishing.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465-471.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, *1*.

Sajadieh, S. et al. (2026). The AI Index 2026 Annual Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, https://doi.org/10.48550/arXiv.2606.15708

Santayana, G. (1925) *Dialogues in Limbo*. London: Constable and Co.

Schmidt-Petri, C., Schefczyk, M., & Osburg, L. (2022). Who Authored On Liberty? Stylometric Evidence on Harriet Taylor Mill's Contribution. *Utilitas*, *34*(2), 120-138.

Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2024). Creating a large language model of a philosopher. Mind & Language, 39(2), 237-259.

Stich, S. (2015). What Is Experimental Philosophy? *Annals of the Japan Association for Philosophy of Science*, 23: 21–31.

Stuart, M. T., Fehige, Y., & Brown, J. R. (2017). Thought experiments: state of the art. *The Routledge companion to thought experiments*, 1-28.

Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition*, *103*(1), 107-119.

Votsis, I. (2011). Saving the Intuitions: Polylithic Reference. *Synthese*, vol. 180(2): 121-137.

Votsis, I. (2024). A neuro-symbolic approach to the logic of scientific discovery. In *International conference on Model-Based Reasoning* (pp. 306-330). Cham: Springer Nature Switzerland.

Votsis, I., Freeborn, D., Ball, B. & Norelli, M. F. (in preparation) 'Reverse Engineering Prompts: Human-to-Human and Human-to-Chatbot Question Answering'.

Warner, J. (2024). Faculty Must Protect Their Labor from AI Replacement. *Inside Higher Education*: December 11, 2026.

Williamson, T. (2000). *Knowledge and its Limits*. New York: Oxford University Press.

Williamson, T. (2015). *Tetralogue: I'm right, you're wrong*. Oxford University Press.

Wittgenstein, L. (1953) *Philosophical Investigations*, G.E.M. Anscombe and R. Rhees (eds.), G.E.M. Anscombe (trans.), Oxford: Blackwell.

Zangari, L., Greco, C. M., Picca, D., & Tagarelli, A. (2025). A survey on moral foundation theory and pre-trained language models: Current advances and challenges. *AI & SOCIETY*, 40(6), 4973-4998.

Zollman, K. J. (2015). Modeling the social consequences of testimonial norms. *Philosophical Studies*, *172*(9), 2371-2383.