Computational Analysis for Philosophical Education: A Case Study in AI Ethics

Brian Ball, Alex Cline, David Freeborn, Alice Helliwell, and Kevin Loi-Heng (Forthcoming)
Philosophical Education.

Abstract

This paper explores what computational methodologies can tell us about philosophical education particularly in the context of AI Ethics. Taking the readings on our AI Ethics and Responsible AI syllabi as a corpus of AI ethics literature, we conduct an analysis of the content of these courses through a variety of methods: word frequency analysis, TF-IDF scoring, document vectorization via SciBERT, clustering via K-means, and topic modelling using Latent Dirichlet Allocation (LDA). We reflect on the findings of these analyses, and more broadly on what computational approaches can offer to the practice of philosophical education. Finally, we compare our approach to previous computational approaches in philosophy, and more broadly in the digital humanities. This project offers a proof-of-concept for how contemporary NLP techniques can be used to support philosophical pedagogy: not only to reflect critically on what we teach, but to discover new materials, explore conceptual gaps, and make our courses more accessible to students from a range of disciplinary backgrounds.

Key Words: AI Ethics, Philosophy Education, Computational Philosophy

1 Introduction

What can computational methods - particularly artificial intelligence (AI) - tell us about AI ethics education? In this paper, we apply computational approaches to interrogate our AI ethics courses. As philosophers working in the philosophy of AI, we are interested in what computational methods can add to philosophical studies, and vice versa.

In our philosophy programmes [redacted for peer review] AI ethics education forms an integral part of our teaching. Our students have a wide range of backgrounds, though of course many have been trained in either philosophy or computer/data science. Our aim is to provide philosophical and computational education simultaneously, to equip students with the skills they need to responsibly engage with AI technology. Given this ethos, we have decided to turn use of computational methodologies on our own practice, by investigating some of the philosophy courses on these programmes. Our aim is to gain insight into our pedagogical approach and to develop a project which we can (hopefully) share with our students. In order to test our thought that computational tools can be useful for pedagogical and philosophical goals, we have conducted a computational analysis of the texts we set for students across two courses in AI Ethics [Course titles redacted for anonymous review]. We have curated these papers over several years, and after completing both courses, we want our students to have covered a variety of classic and current topics in AI ethics and responsible AI. Having gathered the recommended texts for these courses, we utilised some standard Python-based natural language processing (NLP) techniques to analyse our corpus of texts.

In this paper, we explain our methodology and discuss the results of our analysis. We begin (section 2) with a description of the dataset, consisting of the reading materials assigned in two of our advanced (advanced undergraduate, MA and MSc level) philosophy courses on AI and data ethics, and explain how we prepared

the texts for computational analysis. In section 3, we discuss the ethical considerations for this project. In section 4, we describe the natural language processing (NLP) techniques we used to explore this corpus: from relatively simple tools such as word frequency analysis and TF-IDF scoring, to more complex machine learning approaches including document vectorization via SciBERT, clustering via K-means, and topic modelling using Latent Dirichlet Allocation (LDA). Each of these methods offers a different lens through which to understand the themes of our syllabi. Word frequency and TF-IDF give us a surface-level, yet still informative, comparative view. SciBERT vectorization and clustering allow us to explore semantic relationships within the corpus. Topic modelling, finally, enables us to identify and interpret latent themes running throughout this body of literature.

Finally, in section 5, we discuss the broader implications of our approach, both for AI ethics education and for philosophy more generally. As philosophers working on AI, we see this project as a two-way exchange: using computational tools to enhance philosophy teaching and using philosophy to reflect critically on the use of such tools. We situate our work in the context of digital humanities, noting that while computational methods have been widely used in literature, history, and linguistics, they remain relatively underexplored in philosophy. This project offers a proof-of-concept for how contemporary NLP techniques can be used to support philosophical pedagogy: not only to reflect critically on what we teach, but to discover new materials, explore conceptual gaps, and make our courses more accessible to students from a range of disciplinary backgrounds. We conclude (section 6) with a call for further work in computational philosophy and philosophical pedagogy — and outline our plans for future analysis and engagement with students as collaborators in this ongoing exploration.

2 Data Set

The dataset utilised in this research consists of the required and supplementary readings assigned in two upper-level philosophy courses we teach: [precise titles redacted for peer review]. These are graduate-level or advanced undergraduate courses aimed at students from diverse disciplinary backgrounds, including philosophy, computer science, and the social sciences, as well as many graduate students with experience in industry. As instructors, we have curated the readings to offer both foundational and contemporary perspectives in the broad field of AI and data ethics. The goal is to introduce students to a wide range of normative concerns and philosophical methods, while also equipping them with the analytical tools to evaluate real-world technologies, applications and policies. After completing the two courses, students should have established knowledge of essential topics in responsible AI and AI ethics, as well as the necessary skills to engage in normative discussions on emerging advances in AI.

The selected readings include a mix of philosophy papers, technical and policy-oriented research, and interdisciplinary contributions from fields such as computer science, law, economics and education. Authors in the corpus range from prominent philosophers to computer scientists discussing algorithmic bias, as well as economists, legal scholars writing on data privacy and AI regulation, and some technical practitioners of AI.

Together, the two courses span 22 weeks of teaching and 17 distinct thematic topics. Topics in course 1 [redacted] include:

- What is AI and Data Ethics?
- Autonomous AI and Responsibility
- Artificial Moral Agency

- Personhood and Robot Rights
- Algorithmic Bias and Fairness
- Safe AI (including Black Boxes, Transparency, and Explainability)
- Data, Democracy, and Misinformation
- Privacy and GDPR
- Superintelligence and the Control Problem
- Regulation
- Value Sensitive Design

Topics in course 2 [redacted] include:

- What is Responsible AI?;
- AI and Work;
- AI and the Creative Industries;
- AI and Education;
- AI and Human Interaction;
- AI and Sustainability.

From these topics, we collected the full set of assigned readings, resulting in a corpus of 184 distinct texts. These included journal articles, book chapters, and reports. From a technical perspective, we treated each reading as a single document in our corpus. The documents were compiled in plain text format.

Before we could move into computational analysis, we focused on preparing the textual data. Clean and standardized text is essential to ensure that any patterns we uncovered would be meaningful and as free from noise as possible. This step sets the foundation for later stages of the project, including vectorization and clustering. Without a careful cleaning and normalization process, later stages like similarity measurement or topic modelling become vulnerable to distortion by irrelevant or redundant information. The SpaCy model was a useful, lightweight tool for our natural language processing tasks. The tool allowed us to tokenize the text into words and sentences, lemmatize words to their base forms, and remove punctuation and irrelevant characters. The aim here was to ensure that related terms — such as “machines” and “machine” for instance — would be treated consistently.

3 Ethics

Several ethical issues were considered when conducting this analysis. We did not use any human subjects, and also did not utilise any personal information in our analysis, so human subject considerations were not applicable. The authors of the courses under analysis are all part of the project team and granted permission for their syllabi to be used for this analysis.

As we are interested explicitly in AI ethics in this paper, we also considered the ethics of the use of texts for analysis by AI. Whilst the texts in our corpus were all available online, and particularly for educational purposes, we have not made this corpus openly accessible in order to ensure we do not breach copyright protections. As we are utilising AI to analyse our corpus of texts, we were also particularly aware of current debates in intellectual property and AI.[1] There is a growing debate around training data, reproduction, and

[1] Our use of these texts falls under the scope of academic research and teaching, and we believe it is justified under principles of fair dealing, particularly given the non-commercial and scholarly nature of the work, in compliance with

attribution in the context of generative AI. However, the tools used in this project, including word frequency counters, TF-IDF models, SciBERT embeddings, and topic modelling, are all predictive rather than generative. As such, none of these methods produced new textual output derived from the source material; rather, we deployed these tools to extract patterns and representations from the existing dataset, in ways that are standard in computational linguistics and the digital humanities.

Finally, while student engagement is an important motivation for this work, no student data was utilised for this research.

4 Analysis

4.1 Word-frequency Analysis

The first analytic tool we turned on our corpus of texts was a word frequency counter. This simple computational technique counts the number of times a word appears in a document, or collection of documents. This allowed us to identify the words that appear most frequently in our collection of papers, and produce the word cloud, where the most frequently used words appear largest in size, shown in image 1 below.

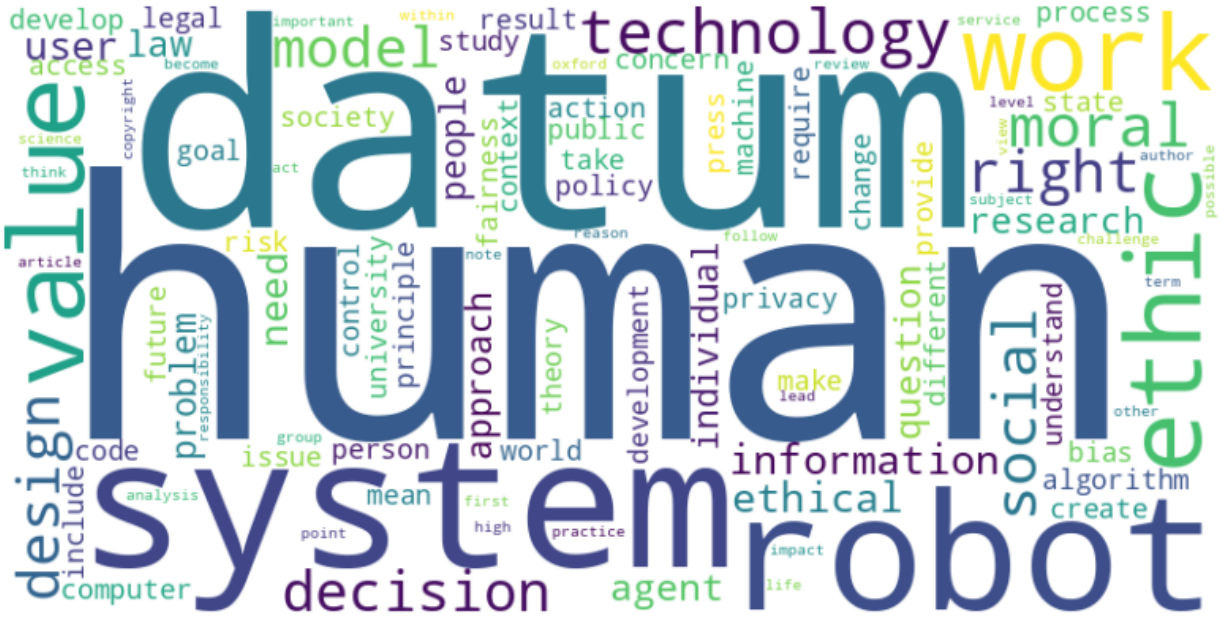


Image 1: word cloud of frequently appearing words in the corpus.

We found that across our two courses, the highest frequency unique word used was ‘human’. As our courses are primarily focussed on technology and ethics, this was perhaps somewhat surprising. However, it is likely that authors in our corpus are discussing humans in contrast with data (second most common) and machines (eighth most common) that are their explicit focuses. Words such as ‘work’, ‘right’ ‘value’ ‘bias’ and ‘ethic’ are to be expected, given the topics in our courses. However, words such as ‘press’, ‘social’, ‘individual’, ’public’ and ‘state’ do not obviously correspond to particular topics and seem to highlight the social and

Section 29 of the UK Copyright, Designs and Patents Act 1988. (UK CDPA 1988, Section 29A — “Copies for text and data analysis for non-commercial research”.)

societal focus of the courses.

Term frequency itself has limited utility for telling us about unique features of a corpus of texts. It could be, for example, that (contrary to the conjecture above) 'human' is something that comes up in philosophical works in general. To find out more about the unique features of this body of texts, we conducted another analysis.

4.2 TF-IDF

To further examine whether our conjecture regarding word frequency was plausible, we decided to analyse word-frequency further. We ran another measure on the corpus: a TF-IDF (Term Frequency–Inverse Document Frequency) (Spärck Jones, 1972). This NLP technique is typically used to evaluate the relative importance of a word in a document compared to its importance in the corpus as a whole. Rather than simply counting the frequency of use for each word, a TF-IDF can show which words are more common in our AI Ethics corpus compared to a larger, or alternative, corpus of texts.[2]

Top 10 words: AI Ethics Canon	**Top 10 words: Wittgenstein Corpus**
human	philosophy
ethic	Wittgenstein
moral	philosophical
robot	language
data	theory
system	political
technology	social
design	review
agent	science
develop	knowledge

Table 1: Top 10 words in each of the datasets

Of course, in this case we were not just interested here in individual papers, but the body of works as a

[2] Term Frequency (TF) measures how often a term appears in a document relative to the total number of terms in that document. For a given term t_j, the term frequency is defined as $TF_j = t_j / \Sigma t_i$, where t_j is the number of times term j appears in the document, and Σt_i is the total number of terms in the document. However, words that appear frequently across the entire corpus may be less informative.To account for this, inverse term frequency (IDF) is defined by, $IDF_j = \log(N / (1 + n_j))$, where N is the total number of documents in the corpus, and n_j is the number of documents in which term t_j appears. The "+1" in the denominator avoids division by zero. We then define the TF-IDF score for term t_j as the product, $TF\text{-}IDF_j = TF_j \times IDF_j$.

whole. In order to complete a TF-IDF measure then, we required a contrasting corpus of texts. [Redacted for peer review] we had a 'Wittgenstein Corpus' available; a body of papers (accessed through JSTOR) discussing the work of Wittgenstein. This corpus is comprised of 64,000 total documents, was made on *Constellate* (from Ithaka) with their dataset builder from papers on JSTOR.[3]

When we compare these two analyses, we start to see the relative importance of these terms in the text. 'human', for example, is not just the most frequent unique word, but it is particularly important in the AI ethics papers compared to works discussing Wittgenstein. 'Wittgenstein' is the second most important word in the Wittgenstein papers (a comforting sign that our analysis was working). Furthermore, in the Wittgenstein corpus, 'philosophy' and 'philosophical' are particularly prevalent. This may reflect the meta philosophical nature of Wittgenstein's work (and thus discussions of his work) but may also reflect the relative lack of importance of 'philosophy' in the AI ethics corpus, which spans more disciplines (such as law, computer science and engineering).

4.3 Using AI: Vector Representations and Cosine Similarity

Few nowadays would consider the NLP techniques we have discussed so far to involve AI: in particular, the computational methods employed operate directly on textual data, here the full papers from our two course reading lists. Since research papers are written in natural language, they need to be converted into a numerical format that a computer can read and interpret if contemporary AI techniques are to be deployed on them. We did this using SciBERT (Beltagy, Lo & Cohan, 2019), a transformer model pre-trained on scientific texts based on the BERT model (Devlin et al. 2018).[4] SciBERT converts each document into a high-dimensional vector — essentially a mathematical "fingerprint" that captures the semantic content of the text. This allowed us to compare texts not by the words they contain directly, but by their learned representations: encodings that capture patterns of semantic meaning based on usage and context across the corpus (Bengio et al. 2003).

It is helpful to contrast this with another common technique in digital humanities, which is to make use of Word2Vec (Mikolov et al. 2013a; Mikolov et al. 2013b). Unlike SciBERT, which creates a single vector for an entire document, Word2Vec assigns vectors to individual words. A model is trained (actually a number of them) on a corpus, and this associates a vector - not with each document, as in our approach, but - with each word. The vector in question is used for next word prediction: that is, the algorithm aims to associate a vector with each word that determines probabilities for the other words in the vocabulary that they occur next (in the corpus). Accordingly, each word's location in the vector space represents its usage (or distribution), within the corpus (i.e. its associations with other words). This vindicates Frith's dictum, 'You shall know a word by the company it keeps' (Frith 1968) - and it allows us to compare e.g. the conceptualizations of words across corpuses.

---

[3] JSTOR Dataset ID: 77934734-096e-6982-c1de-af09599cd73e. Wittgenstein about Philosophy - Applied philosophy, Philosophy - Axiology, Philosophy - Epistemology, Philosophy - Logic, Philosophy - Metaphilosophy, Philosophy - Metaphysics limited to document type(s) book, article from 1900 - 2023

[4] SciBERT is a pre-trained language model based on the BERT (Bidirectional Encoder Representations from Transformers) architecture, specifically trained on scientific texts. In essence, this transforms raw text into numerical representations through a process known as contextual embedding, i.e. generating a vector for each token, based not just on the word itself, but on the surrounding words in both directions. Through sufficient training on a large sample, the model learns which words are most relevant to each other in context, even when those relationships are fairly weak, or the words are separated by long spans of text. For our analysis, we used the pooled output from SciBERT to produce a single vector representation for each document. This vector can be understood as a dense, high-dimensional summary of the document's semantic context.

After generating a vector for each document in our corpus using SciBERT, we computed pairwise cosine similarity scores between them.[5] A similarity score of **1** indicates highly similar documents (identical in vector space), while a score near **0** indicates very different content. This allowed us to measure semantic similarity between papers, providing a foundation for a clustering analysis (see below).

4.4 Using AI: Clustering Papers into Meaningful Groups

We were also interested in drawing out where papers in our canon were grouped together around different subjects and themes. To examine this, we utilised a couple of methods. First, we applied K-Means clustering, an unsupervised machine learning technique that groups papers into clusters based on their similarity (Steinhaus 1957; MacQueen 1967; Jain & Dubes 1988; Pedregosa et al. 2011). It works on unlabelled data (that is data without defined categories or groups). The algorithm first randomly selects central points called centroids then uses algorithms to automatically find common themes and structures in the data. We repeated the clustering with different k values to find different groupings. By experimenting with different k values we determined the best number of clusters. For this we used techniques like the Elbow Method and Silhouette Score to find a suitable number given the trade-off between better representing the data and using more clusters. We picked six clusters to move forwards.

We tested a range of values for k, the number of clusters, varying the number of clusters from 1 to 32, specifically testing k in [1,2,3,4,5,6,8,12,16,20,24,28,32]. For each clustering solution, we evaluated the results using the silhouette score (see image 2),[6] a standard metric for assessing the quality of clustering (Rousseeuw 1987). The silhouette score captures both cohesion (how close each document is to the other documents in its cluster) and separation (how far it is from documents in other clusters). Scores range from -1 to 1, with higher values indicating more well-defined and internally coherent clusters.

After identifying candidate values of k that produced relatively high silhouette scores, we further examined the resulting clusters to evaluate their interpretability. This involved identifying central documents — those that were closest to the centroid of their cluster — as well as outlier documents that were located on the periphery of a cluster or between two clusters.

[5] Cosine similarity measures the angle between two vectors in a high-dimensional space, given by their normalised dot product. The idea is simple: if two documents are represented by vectors that point in the same direction, they are semantically similar; if the vectors are orthogonal, then they are unrelated. Unlike the Euclidean distance, which measures how far apart two points are, cosine similarity focuses on the orientation of the vectors rather than their magnitude.

[6] The silhouette score for a given document is calculated as (b - a) / max(a, b), where a is the average distance to other points in the same cluster (i.e. intra-cluster cohesion), and b is the average distance to points in the nearest neighbouring cluster (i.e. inter-cluster separation).

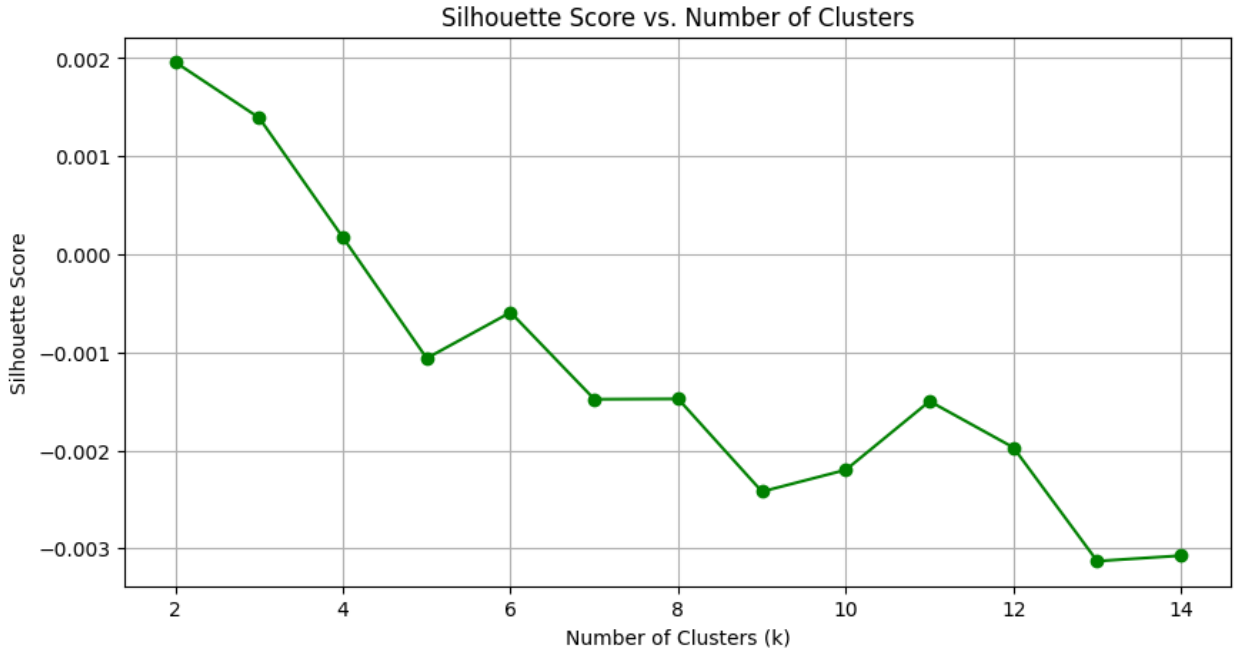
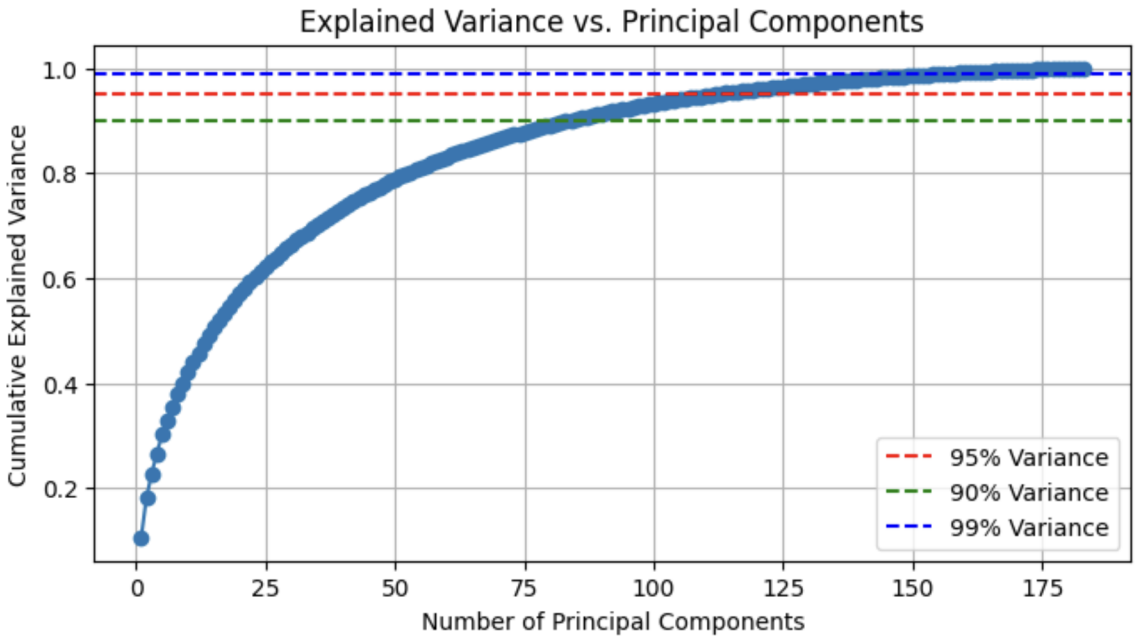


Image 2: Silhouette Score vs. number of clusters

The K-Means clustering algorithm is known to struggle with very high-dimensional data.[7] Since the SciBERT embeddings we used to represent each document exist in a fairly high, 768-dimensional space, we applied a dimensionality reduction technique to make the data more tractable for clustering. To do this, we used Principal Component Analysis (PCA), a linear algebra-based method that transforms the original high-dimensional data into a lower-dimensional space while preserving as much of the data's variance as possible. However, there is necessarily a trade-off between compressing the data, and preserving the salient structural features. PCA works by identifying the orthogonal directions (called 'principal components') along which the data varies the most and projecting the data onto a subset of those directions.[8] In our analysis, we chose to select the number of components such that 95% of the total variance in the original data was preserved (see image 3). This corresponded to 112 principal components, which we used as the input space for the K-Means clustering. The data is shown in image 4, classified into different numbers of clusters and then projected onto just two dimensions for visibility.

[7] This is an example of the so-called 'curse of dimensionality'. Distance metrics, as used for K-Means clustering, become less informative as the number of dimensions increases. In such spaces, all points tend to become approximately equidistant from one another, making it difficult for the algorithm to identify meaningful groupings. Additionally, high-dimensional data tends to be sparse, which further reduces the effectiveness of clustering algorithms that assume dense, well-separated regions.

[8] More formally, PCA finds a new set of orthogonal axes — linear combinations of the original dimensions — ordered by the amount of variance in the data they explain. The first principal component captures the largest possible variance, the second captures the largest variance orthogonal to the first, and so on. By retaining only the top N components, we reduce the dimensionality of the data while maintaining the majority of its informational structure (Joliffe 2002).




```
Number of components to retain 90% of variance: 84
Number of components to retain 95% of variance: 112
Number of components to retain 99% of variance: 157
```

Image 3: Explained variance vs. Principal components.

To ensure that the clustering results were meaningful, we checked whether each paper had the highest similarity to the average of its assigned cluster. The fact that 100% of papers were most similar to their own cluster's average reassured us that the model was making reasonable groupings.[9] In order to visualise these clusters, we needed to conduct further processing on this data, again using PCA, to reduce the clusters to two dimensions.

[9] While this result is not guaranteed by the clustering algorithm, it provided additional reassurance that the groupings reflected real semantic structure in the data.

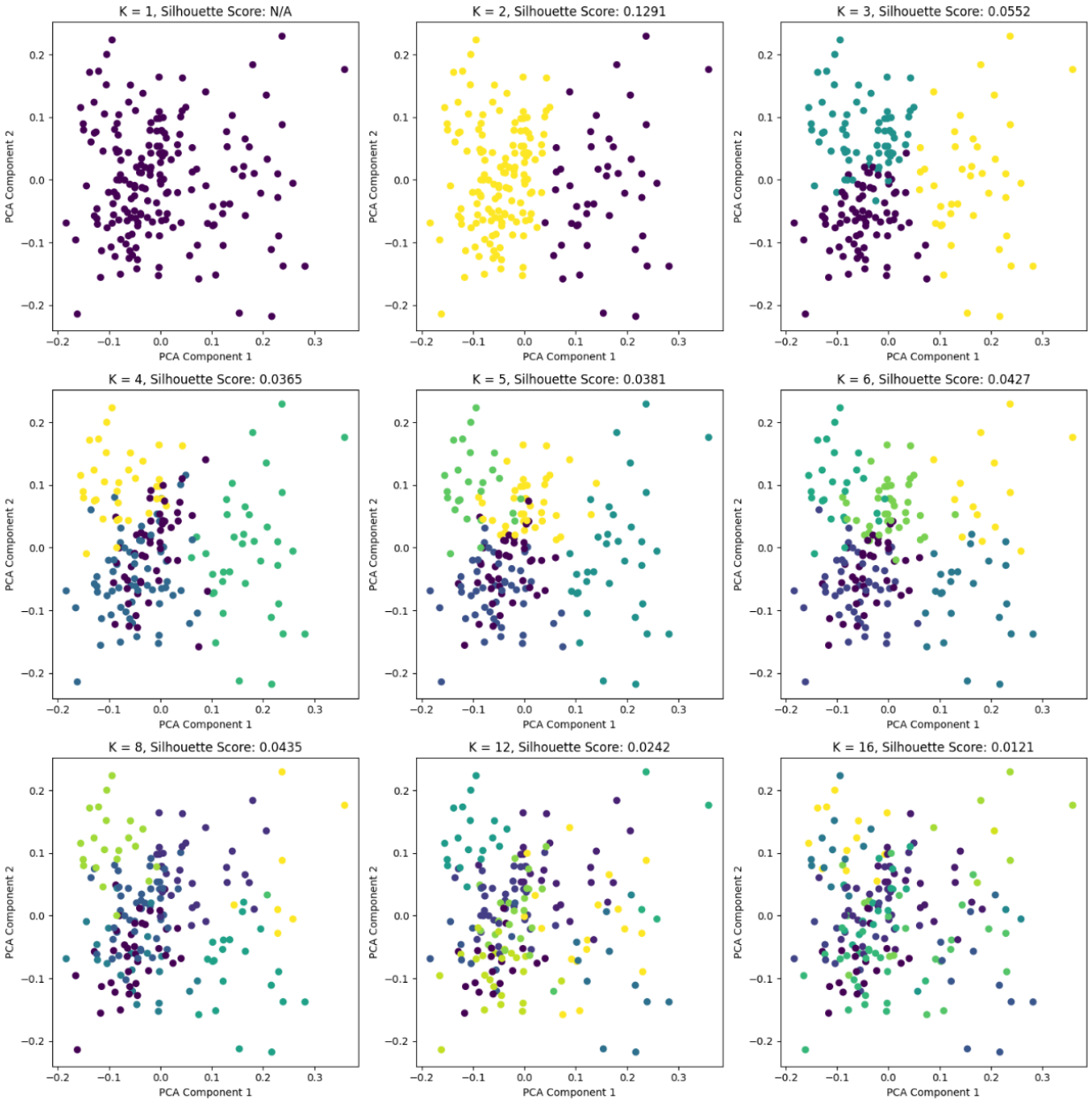


Image 4: K-means clustering with principal component analysis showing the division of papers depending on different numbers of clusters. The dots represent the papers in the canon, with colours representing the clusters to which they belong.

When we looked at which papers fell in each cluster, however, we had a hard time interpreting these clusters. We could not clearly determine which topic/s in AI ethics were key for each cluster. This was likely due to the high dimensionality, and the small number of papers included in our analysis. We are reminded that contemporary AI relies on *big* data, and thus a larger dataset may be necessary to yield interpretable results with this analysis method. We therefore tried an alternative method for grouping the papers in our canon.

4.5 Using AI: LDA Topic Analysis

We next used the Latent Dirichlet Allocation (LDA) method to examine the canon, to see if the paper groupings produced made more sense to us. Like K-Means clustering, LDA is an unsupervised machine learning approach (Prichard, Stephens & Donnelly 2000; Falush, Stephens & Pritchard 2003; Blei, Ng & Jordan 2003). However, unlike K-means, we can use LDA to gather papers under topics, and to then produce a list of words for each topic, making it more interpretable.

LDA is a soft clustering method, which models probability distributions over words and documents. When we use LDA to analyse papers, it treats each paper as an unstructured 'bag of words'- i.e. it does not consider the position of each word in the paper (unlike SciBERT). LDA builds a model of the whole corpus, producing a conditional joint probability distribution of a topic given a word, or a topic given a collection of words (i.e. a paper). This means that LDA tries to identify distinct topics by finding correlations between words. Frequent co-occurrence of words suggests they are related in a topic, whereas non-co-occurrence of words suggests they are not related in a topic.[10]

Our output from LDA is a series of probabilities. For each paper (collection of words) we get a probability that it falls in each topic (here, 6 possible topics). A paper is therefore not just assigned to one topic - instead, it can have a high probability of concerning multiple topics. This may be for good reason - for example, an overview paper might end up having a high probability of concerning e.g. 'privacy' 'AI design' and 'robot agency' (etc). From examining the topics uncovered in this manner, we felt like we could make some sense of them. We identified the broad themes of each topic as follows:

Topic clusters:

0: Social, social media, gender, culture

1: Superintelligence

2: Applied issues such as sustainability, health, and the arts

3: Robots, personhood, and artificial agency

4: Design, responsibility

5: Privacy and risk

To prepare the corpus for topic modelling, the cleaned AI Ethics texts were first transformed into a document-term matrix using a bag-of-words approach. This matrix represents each document as a vector of word counts, capturing the frequency of the 1,000 most common words across the entire corpus (lower frequency words were not included for reasons of computational tractability).

[10] LDA builds a Bayesian probabilistic model of a corpus. It assumes that each document is a mixture of latent topics, and that each topic is characterized by a distribution over words. Formally, LDA posits the following generative process: for each document, a distribution over topics is drawn from a Dirichlet prior; then, for each word in the document, a topic is sampled from that distribution, and a word is sampled from the corresponding topic's word distribution (also drawn from a different Dirichlet prior). The model infers the topic and word distributions that best explain the observed word co-occurrence patterns in the corpus

We then trained the LDA model, specifying that it should extract six topics from the corpus. This decision was informed by the earlier steps in our analysis. In particular, when applying PCA followed by K-Means clustering, we observed signs of natural groupings in the data. Experimentation with different values of k, combined with inspection of Silhouette Scores, suggested that a range of 5 to 8 clusters produced reasonably coherent and interpretable partitions without over-fragmenting the data. Selecting six topics allowed us to strike a balance between granularity and conceptual clarity. After fitting the LDA model, each document was assigned a probability distribution over the six topics. To interpret the model, we identified each document's most probable topic — that is, the topic to which it had the highest posterior probability of belonging. This provided a way of associating each paper with a dominant thematic group, based on its characteristic patterns of word usage. We also identified the number of papers in common between topics (image 5).

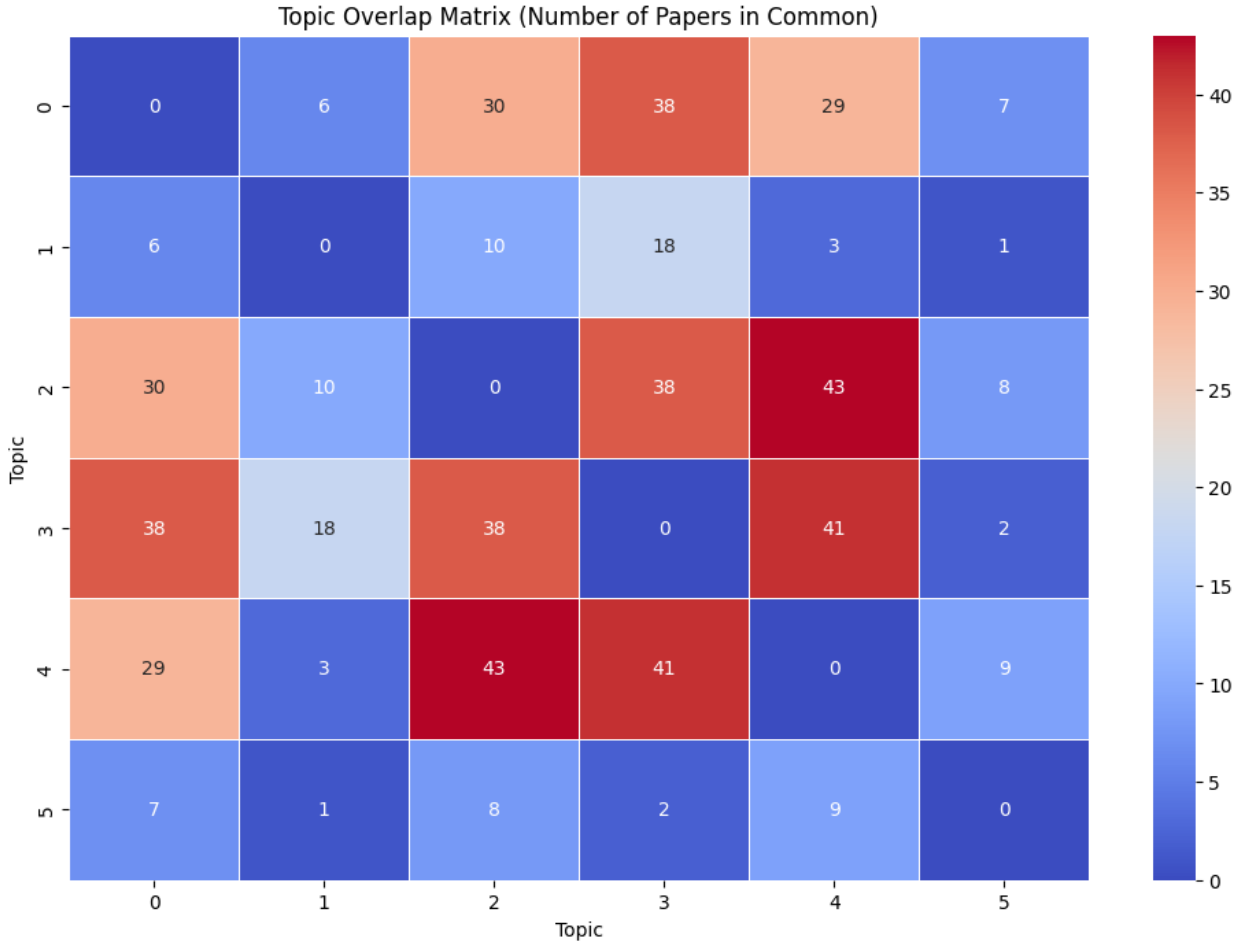


Image 5: Topic overlaps, showing the number of papers which fell in the overlap of each of the six identified themes.

These topics certainly seemed to us to have some internal unity (as indicated), but they could also be seen not to overlap one another in problematic ways. Looking at the percentage of the papers in one topic (the row in the above table) that overlapped with papers in the other topic (in the columns), we found both that the overlap was not in general too great, and that the overlaps present could also be readily interpreted. For example, 52.9% (18) of the papers on superintelligence (topic 1) could also be viewed as concerned with a

topic involving the notion of artificial agency (topic 3), which is understandable given that ethical concerns around the former appeal to the latter; moreover, looking at the column corresponding to superintelligence, we see that it is entirely blue, meaning that none of the other topics overlapped much with it - and indeed, our impression from working within the field is that this topic does, as a matter of sociological fact about the AI ethics community, stand somewhat apart.

5 Discussion

5.1 Reflection and learning

The K-Means clustering, while methodologically sound and internally coherent (as shown by cosine similarity to cluster centroids), ultimately proved difficult to interpret. Although the algorithm grouped texts into clusters based on semantic similarity, we found that the resulting groupings did not consistently align with recognisable course topics or thematic divisions. This may reflect the relatively small size of our corpus, the high dimensionality of the vector space, or the fact that many papers engage with multiple overlapping concerns, making clear separation into exclusive clusters difficult. While the exercise corroborated our preprocessing and embedding pipeline, it may suggest certain limits of hard clustering techniques in the context of philosophical and interdisciplinary content. With this said, it may be that this technique may work more effectively with larger or more varied datasets, or that other dimensional reduction techniques might be needed that better capture the salient structural features of the data, before clustering is applied.

In contrast, the topic modelling using LDA proved to be much more informative. The topics inferred by the model corresponded to intuitively meaningful groupings, such as privacy and risk, robot personhood, or design and responsibility. This method exposed thematic threads that cut across the weekly course topics. Importantly, because LDA provides probabilistic topic distributions, it allowed us to see how individual papers often straddled multiple themes, capturing relations that course structures may obscure. In this sense, LDA may be especially well-suited to philosophical corpora, where overlapping normative, conceptual, and technical concerns are the norm rather than the exception.

Of course, it is important to recognize that no computational tools are methodologically neutral: their meaningful interpretation rests upon assumptions about how the data is structured and what counts as significant. For example, TF-IDF and LDA both treat terms as discrete lexical units, abstracted from their syntactic and argumentative context. On the other hand, SciBERT vectorization is sensitive to local linguistic context but will inevitably encode biases from its architecture and training data. We treated semantic similarity as a linearly decomposable property, geometrically represented by cosine similarity in a high-dimensional vector space. This treats meanings as comparable via vector directions and distances, implying that semantic relationships, such as the distinction between ‘privacy’ and ‘transparency’, can be consistently represented as angular differences across the embedding space. In this sense, even our transformer methods may be insensitive to some contextual subtleties (Ethayarajh, 2019). Likewise, K-means clustering imposes a fixed number of discrete non-overlapping, roughly isotropic clusters, an assumption unlikely to hold in domains with overlapping, intersecting or multifarious concerns. Principal Component Analysis (PCA) assumes that the most meaningful structure in the data lies along orthogonal axes of maximal variance, treating key concepts as essentially uncorrelated (Jolliffe, 2002). The principal component dimensions will not necessarily correspond to conceptual or pedagogical importance. Such

assumptions may be justified as reasonable approximations of the real data or by the practical utility of the methods. However, it is essential to recognize them when drawing conclusions from the results. We contend that these tools are best understood not as offering definitive answers, but as producing artefacts that require philosophical interpretation.

In terms of the pedagogical utility of the approach we have undertaken, we have found that the process has yielded discussion and reflection of our core modules in AI ethics. For future iterations of our courses, we can utilise topic words to help identify new literature in areas that are directly related to our course topics, which may help to diversify our recommendations for students. Particularly notable are the areas of overlap, which could be emphasised in our courses to enhance student understanding of the AI Ethics landscape. The areas where there is little overlap also interestingly suggests that there may be areas of AI Ethics which remain distinct from one another, highlighting potential areas for further exploration (though analysis of a larger corpus would be needed to verify this). In addition, we plan to discuss the results of our analysis with our students, reflecting on the overlapping themes of the courses that go beyond the delineated weeks of the course. For example, the six clear topics we uncovered through LDA did not exactly correspond to our 17 course topics; some were unsurprising (such as agency, personhood and robot rights) however others (such as design and responsibility) fall under different sections of the course. Such insights (for example how responsibility can hinge on design choices) may provide stimulating discussion on our courses. Given their aims, noted above, of simultaneously providing the philosophical and computational education needed for students to engage with the realities of responsible AI, we also expect that it will be valuable to discuss the methodological issues we have encountered along the way - such as the difficulties of using K-means clustering on sparse data distributed in a high-dimensional space. We may also discuss with them the value of AI assistance, as opposed to full automation, as regards our own ongoing course design: for example, we in no way regard the identification, within our data, of fewer topics than were initially conceived by our course leaders as in any way impugning the expert human judgment that went into our course design; rather, we plan to use the AI-generated insights discussed above to supplement our own decision-making in adapting and revising our syllabi in the ways indicated. This, of course, is a point that applies much more broadly, both within applications of AI for philosophical education, and indeed in other domains more generally.

5.2 Computational Analysis for Philosophy

Computing and philosophy have long been intertwined.[11] There are professional bodies dedicated to (aspects of) their intersection, such as the International Association of Computing and Philosophy, as well as the Society for the Philosophy of Artificial Intelligence.[12] And (as noted by Weinberg 2016) there are, of course, some notable examples of excellent - and early - digital resources in philosophy: *the Stanford Encyclopedia of Philosophy* (Zalta & Nodelman created in 1995); the online (and open access) journal *Philosophers' Imprint* (established in 2001); and *PhilPapers* (begun in 2009).[13] Nevertheless, relatively few philosophers have followed the famous suggestion from Leibniz:

> *If controversies were to arise, there would be no more need of disputation between two philosophers than between two accountants. For it would suffice to take their pencils in their hands, to sit down*

[11] As a matter of fact, in our own University, the two disciplines initially sat within the same academic unit, or Faculty.

[12] International Association of Computing and Philosophy, IACAP https://www.iacap.org/. Society for the Philosophy of Artificial Intelligence (PHAI) https://philai.net/.

[13] *The Stanford Encyclopedia of Philosophy* https://plato.stanford.edu/; *Philosophers' Imprint* https://journals.publishing.umich.edu/phimp/; *PhilPapers* https://philpapers.org/

with their slates and say to each other [...]: Let us calculate!

That is, 'philosophers have arguably failed to take full advantage of the opportunities afforded' (Ball et. al 2024, 2) by the computational methods that are both available and widely used in the (other) humanities (disciplines). For example, in one list of 145 academic journals dedicated to the digital humanities, a search for 'philosophy' yields 0 entries (whereas 'humanities' gets 15 hits, 'history' has 4, and 'literature' 2).[14] Nor are there many pertinent results on Google scholar when one searches for 'digital humanities philosophy', 'digital philosophy' or even 'computational philosophy'. This last term has, however, gained some fluency, and there is even an SEP article dedicated to the topic (Grim and Singer, 2024): though that piece is largely concerned with (what has been dubbed) 'simulation as a core philosophical method' (Mayo-Wilson and Zollman, 2021); a quick search of its contents reveals NO mentions of 'natural language processing' (NLP) or 'large language models' (LLMs) - techniques and tools that very widely used in the digital humanities, for both research and teaching purposes… and of course in the pedagogical research we have embarked upon here.

Still, there are some existing digital projects in philosophy, and we shall accordingly devote some (brief) space to their discussion. Many involve data visualizations - for example, *the Philosopher's Web* is a (self styled) 'comprehensive map of all influential relationships in philosophy according to Wikipedia'.[15] It is described in more detail in Jones (2017) and Weinberg (2017), but in brief it shows key figures in philosophy, providing short bios (for some of them), and showing connections (specifically, relations of influence) between them. It does not have a pedagogical focus, but could nevertheless be useful for teaching (perhaps especially the history of) philosophy. Many also involve SEP data.[16] Thus, V*isualizing SEP* does precisely what its name says it will:[17] Stanford Encyclopedia articles are classified (based on the taxonomy developed by the *Internet Philosophy Ontology Projec*t[18]), and links to other articles on the same topic(s) are shown. This might be pedagogically useful for students (or researchers) engaged in a literature search - that is, for those trying to figure out what to read as they begin on a new topic. (Indeed, philosophy teachers might conceivably look to it when constructing a new course.) Finally, *History of Philosophy: Summarized and Visualized* is a hand curated visualization of the positions held by philosophers, and their connections - both supporting and conflicting - with theses espoused by other philosophers (Önduygu 2025).[19] Again, it is not primarily a pedagogical project, but might well have pedagogical uses: in particular, it is potentially useful to students to have the substance of the various philosophers' views articulated, and the semantic, or logical, relations between them displayed (and navigable).[20]

Some projects, like ours, are research-oriented, and even involve natural language processing. For example, Mark Alfano has called (Alfano 2018) for collaborators to engage in a semantic mapping project in

[14] Available at: https://dhjournals.github.io/list/ (Spinaci, Colavizza & Peroni 2020; 2022).
[15] https://kumu.io/GOliveira/philosophers-web#map-b9Ts7W5r
[16] As in *The directed graph of SEP related-entries* (https://mboudour.github.io/2020/05/06/Graph-of-references-among-entries-of-the-Stanford-Encyclopedia-of Philosophy.html), which provides a (someone difficult to see) network representation of the articles in the SEP, and the links between them. The dataset underlying this visualization is no doubt of interest - but we prefer to discuss the alternative example in the main text.
[17] Visualising SEP https://www.visualizingsep.com/#
[18] Internet Philosophy Ontology Project https://www.inphoproject.org/taxonomy. This is a research project funded by the NEH. It is committed to open data - so its code is available, as are the taxonomies generated; and there are research papers on the site describing the approach taken. The project is not primarily pedagogical in focus, and only students with fairly advanced technical skills would be well-placed to engage with it in any detail.
[19] https://www.denizcemonduygu.com/philo/
[20] Other projects in the same spirit as those discussed in this paragraph are touched upon in Weinberg (2014).

philosophy, using texts available from Project Gutenberg - and promises to create freely shareable teaching materials! The digital humanities approach underlying the project is described in another blog post (Alfano 2017): it is not unlike the Word2Vec description given in the main text above, though it relies, perhaps, on a different computational technique.

Amongst projects with a pedagogical focus, some are relatively straightforward: TeachPhilosophy101 for example, is principally a website with materials - including digital resources - that may be useful to teachers of philosophy.[21] Others involve more comprehensive data analysis: for example, Open Syllabus Galaxy maps the most assigned readings across 7m+ course syllabuses.[22] And still others are more targeted: for example, ArgumApp is a pedagogical app concerned quite specifically with argument mapping (Mohler 2020);[23] and The Logic Calculator tests for syntactic well-formedness and semantic validity in the propositional calculus (Votsis, 2019).[24] And some seem mostly designed for fun. For example, Weinberg (2021) highlights Maximilian Noichl's SEP haiku project. This project involves searching the SEP for strings of 17 syllables and then checking whether the word breaks fall in the right places to make a haiku. If so, it makes that haiku. The materials produced could be used by teachers looking to find appropriate tidbits to introduce lectures, or to serve as mnemonics for students.

This is by no means an exhaustive overview of the digital projects that have been pursued in relation to philosophy, or philosophical pedagogy, but it is not entirely unrepresentative in our view. And if we are right about that, it should be clear that what we have done here is quite atypical, at least within philosophy. For we have turned quite heavy-duty computational methods upon our own teaching practice - the syllabi we have created - to see what they reveal about the contents of our courses.

Indeed, we pause to briefly dwell on the novelty of the approach taken here, not only relative to existing practices within philosophy, but even in the context of digital humanities as a whole. Advances in AI have come fast and thick in recent years, bringing disruption across all aspects of society. Digital humanities can hardly be expected to prove an exception - and indeed, some scholars have begun to grapple with the question of how to incorporate advanced NLP techniques into humanities research (Suissa, Elmalech & Zhitomirsky-Geffet 2022; Ehrmanntraut et al 2021; Liu et al. 2024). And yet, to the best of our knowledge, ours is the first attempt within the humanities to use transformer-based vector embeddings of whole documents to provide distant readings for the analysis of a corpus.[25] While this particular method has not yielded deep insights in looking at our relatively modestly sized corpus - and certainly none that can themselves be generalized to e.g. other (individual) syllabus analyses - we anticipate that this pioneering approach, as we develop and refine it further, or at least its ultimate assessment (e.g. through comparison with the older LDA-based technique), will prove valuable well beyond the present context.

In future research we will continue exploring the possibilities of this analytic approach for AI Ethics and philosophical pedagogy. In particular, we plan to analyse a wider corpus of texts in relevant fields. We hope that this will help us gain a better understanding of relevant literature, identify emerging topics as well as literature gaps, and draw on uncovered connections between topics and bodies of work to signpost to our students.

6 Conclusion

[21] See https://www.teachphilosophy101.org/
[22] See https://galaxy.opensyllabus.org/
[23] https://appsolutelyfun.com/argumap.html
[24] See https://votsis.org/logic.html
[25] However, see efforts by Cohan et al. (2020) to adapt similar methods in other fields.

We have demonstrated how computational analysis of readings on philosophical syllabi can yield useful reflections for educators in philosophy. Our dataset consisted of the materials assigned in two of our philosophy courses in the field of AI ethics. We prepared this dataset for analysis, taking into account any ethical concerns with our proposed approach. We implemented several NLP techniques to analyse our corpus. We began with relatively simple approaches (word frequency analysis and TF-IDF) which yielded some noteworthy results, particularly the relative importance of the 'human' in the AI ethics course corpus, and the relative unimportance (compared to the Wittgenstein corpus) of 'philosophy'. Given the nature of these approaches, only limited conclusions could be drawn. We then moved on to more complex NLP approaches including document vectorization via SciBERT, clustering via K-means, and topic modelling using Latent Dirichlet Allocation (LDA). SciBERT vectorization and clustering allowed us to explore semantic relationships within the corpus, however we struggled to draw conclusions from this approach, likely due to the small number of papers in our corpus. In future analyses we plan to use a larger dataset in order to combat this limitation. Topic modelling through LDA enabled us to identify six broad themes in the corpus, which were in some cases different to what we might expect given the topics we set and how these are connected on the course. Finally, we discussed the broader implications of our approach, both for AI ethics education and for philosophy as a discipline. Given the limits of existing work in computational approaches in the field of philosophical research (even in AI ethics) we see an opportunity to harness these approaches for philosophy and philosophical education.

Acknowledgements
We would like to thank Northeastern University's NULab for Digital Humanities and Computational Social Sciences, the Internet Democracy Initiative, and the Ethics Institute, for funding this research. We would also like to thank attendees at the NULab Spring conference 2025, and attendees and organisers of Philosophical Education's webinar series for their helpful feedback.

References

Alfano, Mark (2017). "A Semantic-Network Approach to the History of Philosophy Or, What Does Nietzsche Talk about when He Talks about Emotion?", *Daily Nous* [Blog] 26 July. Available at: https://dailynous.com/2017/07/26/semantic-network-approach-history-philosophy-guest-post mark-alfano/

Alfano, Mark (2018) "Collaborators sought for digital humanities project on the history of philosophy" *Philosophy and Other Thoughts* [blog] 23 June. Available at: https://www.alfanophilosophy.com/blog/2018/6/23/collaborators-sought-for-digital-humanities project-on-the-history-of-philosophy

Ball, B., Koliousis, A., Mohanan, A. and Peacy, M. (2024). "Computational philosophy: reflections on the PolyGraphs project" *Humanit Soc Sci Commun* vol. 11, no. 186. https://doi.org/10.1057/s41599-024-02619-z

Beltagy, Iz, Kyle Lo, and Arman Cohan. (2019) "SciBERT: A pretrained language model for scientific text." *arXiv* preprint arXiv:1903.10676

Blei, David M., Andrew Y. Ng. and Michael I Jordan (2003). "Latent Dirichlet Allocation". Journal of Machine Learning Research. Vol. 3, no. 4–5: pp. 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. (2020) "SPECTER: Document-level representation learning using citation-informed transformers." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2270–2282). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.207

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. (2019) "Bert: Pre-training of deep bidirectional transformers for language understanding." In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1*, pp. 4171-4186. https://arxiv.org/abs/1810.04805

Ehrmanntraut, A., Hagen, T., Konle, L., & Jannidis, F. (2021) Type-and token-based word embeddings in the digital humanities. In *CHR 2021: Computational Humanities Research* Conference (pp. 16-38). Available at: https://ceur-ws.org/Vol-2989/long_paper35.pdf (Accessed July 2025)

Ethayarajh K. (2019). |How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings". *arXiv preprint*. arXiv:1909.00512,

Falush, D. Stephens, M., and J. K. Pritchard (2003). "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies". *Genetics.* vol 164, no. 4, pp. 1567–1587. doi:10.1093/genetics/164.4.1567.

Firth, J. R. (1968) "Chapter 11: A synopsis of linguistic theory" In *Selected Papers of J. R. Firth 1952-59*, Edited by F. R. Palmer, Reprinted from: Studies in linguistic analysis (Special volume of the Philological Society, Oxford, 1957, 1-31), Indiana University Press, Bloomington, Indiana.

Grim, Patrick and Daniel Singer, (2024) "Computational Philosophy", In *The Stanford Encyclopedia of Philosophy* (Summer 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), Available at: https://plato.stanford.edu/archives/sum2024/entries/computational-philosophy/

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer.

Jones, J (2017) "The Philosophers Web" Open Culture, 20 October. Available at: https://www.openculture.com/2017/10/the-philosophers-web.html

Liu, C., Wang, D., Zhao, Z., Hu, D., Wu, M., Lin, L., Liu, J. Zhang, H., Shen, S., Li, B. and Zhao, L. (2024) "SikuGPT: a generative pre-trained model for intelligent information processing of ancient texts from the perspective of digital humanities." *ACM Journal on Computing and Cultural Heritage*, 17(4), 1-17. https://doi.org/10.1145/3676969

Mayo-Wilson, C., and K. J. S. Zollman (2021) "The computational philosophy: simulation as a core philosophical method". *Synthese,* 199, pp. 3647–3673. https://doi.org/10.1007/s11229-020-02950-3

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, (2013a) "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. (2013b) "Distributed representations of words and phrases and their compositionality." *Advances in neural information*

processing systems 26 (2013). https://arxiv.org/abs/1301.3781

Mohler, Chad (2020) "From Maps to Apps: Introducing Students to Argument-Mapping in the Physical and Digital Realms", *Daily Nous* [Blog] 25 November. Available at: https://dailynous.com/2020/11/25/maps-apps-introducing-students-argument-mapping-guest-post/

Önduygu, Deniz Cem (2025) "New Force-Directed Graph with Philosophers as Nodes" *Deniz Cem Önduygu* [Blog] January 29, 2025. Available at: https://www.denizcemonduygu.com/philo/new force-directed-graph-with-philosophers-as-nodes/

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research* 12 (2011): 2825-2830.

Pritchard, J. K., M. Stephens, and P. Donnelly (2000). "Inference of population structure using multilocus genotype data". *Genetics.* 155 (2): pp. 945–959. doi:10.1093/genetics/155.2.945

Rousseeuw, Peter J. (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Computational and Applied Mathematics*. 20: 53–65. doi:10.1016/0377-0427(87)90125-7

Suissa, O., Elmalech, A., and Zhitomirsky-Geffet, M. (2022). "Text analysis using deep neural networks in digital humanities and information science." *Journal of the Association for Information Science and Technology,* 73(2), 268-287. https://doi.org/10.1002/asi.24544

Spärck Jones, K. (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval" *Journal of Documentation.* 28 (1): 11–21. https://doi.org/10.1108/eb026526

Spinaci, G., G. Colavizza and S. Peroni (2020). "Preliminary results on mapping digital humanities Research", *Proceedings of L'Associazione per l'Informatica Umanistica e La Cultura Digitale*, pp. 246–252, Available at: https://aiucd2020.unicatt.it/aiucd-Spinaci_et_al.pdf.

Spinaci, G., G. Colavizza, and S. Peroni. (2022). "A map of Digital Humanities research across bibliographic data sources", *Digital Scholarship in the Humanities* https://doi.org/10.1093/llc/fqac016.

Weinberg, Justin (2014). "Graphing the History of Philosophical Influences". *Daily Nous* [Blog], 21 April. Available at: https://dailynous.com/2014/04/21/graphing-the-history-of-philosophical-influences/

Weinberg, Justin (2016) "Digital Humanities In Philosophy: What's Helpful & What's Hype?" *Daily Nous* [Blog] May 24. Available at: https://dailynous.com/2016/05/24/digital-humanities-in-philosophy whats-helpful-whats-hype/

Weinberg, Justin (2017) "A Visualization of Influence in the History of Philosophy". *Daily Nous* [Blog] 11 January. Available at: https://dailynous.com/2017/01/11/visualization-influence-history philosophy/

Weinberg, Justin (2021) "Making Haiku and Art from the SEP" *Daily Nous* [Blog], 31 August. Available at: https://dailynous.com/2021/08/31/making-haiku-art-sep/

Votsis, I. (2019). Logic calculator. Retrieved July 17, 2025, from https://votsis.org/logic.html

Zalta, E. and Nodelman U. (eds) (no date) *Stanford Encyclopedia of Philosophy*. Available at: https://plato.stanford.edu/ ISSN 1095-5054