


Application

Advanced Graduate Workshop in Computational Social Science

Maximilian Noichl 

April 2025

Contact Information

Maximilian Noichl, MA
Promovendus
Institute for Theoretical Philosophy
Department for Philosophy and Religious Studies
Utrecht University
m.a.noichl@uu.nl

Project Proposal: The cognitive dynamics of philosophy

How does intellectual progress happen? Is it that we make our arguments watertight until they can hold no more, yield, and give way to new positions, or is it that new ways of imagining open up channels that restructure our collective mental landscapes?

This question about the interplay of the rigid and the intuitive cognitive tools at our disposal is what I want to investigate during this year's AGWCSS. A main avenue for investigation is the case of philosophical discourse, which builds more-or-less logical arguments in tandem with articulating and refining intuitions, and provides a terrific test-bed for the core question of the relationship between argument making and imagination.

To study this in a quantitative fashion, I am drawing on the novel abilities of Large Language Models (LLMs) to make expert human judgements on text. While previous studies in, for example, science-of-science, have relied upon simple tracers of ideas via word frequencies (*e.g.*, topic modeling), LLMs can now, with a certain amount of prompt engineering, answer far more sophisticated questions that previously required organizing large numbers of human coders.

This project, which I am leading, with Prof. Simon DeDeo (SFI External Faculty) serving as mentor, and which I hope to have ready for submission by the end of the summer, is targeted towards a general audience journal such as *Proceedings of the National Academy of Sciences* or *Journal of the Royal Society Interface*. It leverages these new Generative AI powers to study, in the philosophical record, the

co-evolution of, on the one hand, philosophical *positions* (e.g., the position “scientific instrumentalism”, in the philosophy of science, that says that scientific ideas derive their value pragmatically from their ability to predict and intervene, as opposed to describing what’s “real”), and, on the other hand, philosophical *examples* (e.g., the example of how the kinetic theory of gases, which is often used to illustrate the above distinction). Philosophical papers, generically, contain both, enabling us to construct bipartite networks of association across the historical record.

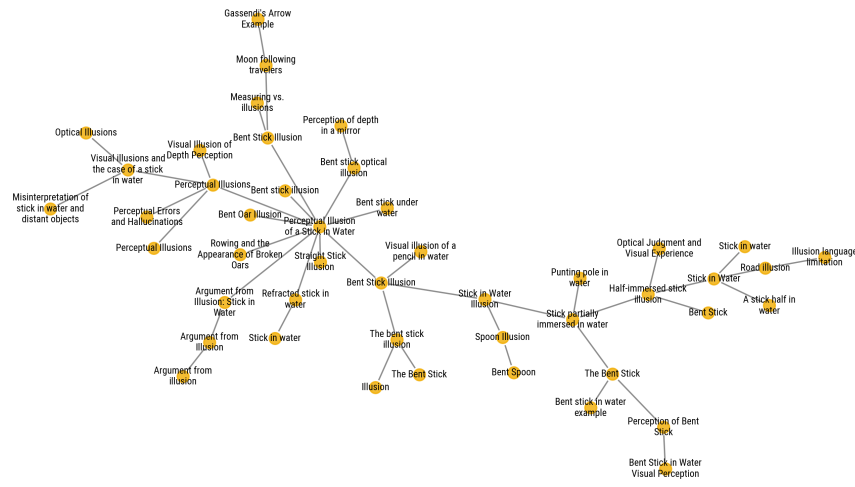


Figure 1: One philosophical example after the merging step. Each node is an instance of use of the example of a spoon/stick half immersed in water, which appears bent due to refraction. This example is commonly used in philosophy to explain the fallibility of visual perception. Nodes are linked if the language model judged them to be the same example.

The data that will underlie this project comes from a sample drawn from all philosophy papers archived on JSTOR, evenly distributed across the period from 1950 to 2015. I thus survey a broad range of recent English-language philosophy.

Given this dataset, I prompt an LLM (GPT-4o) to extract the core structures of philosophical reasoning I am interested in, namely arguments, positions, philosophical views (short: positions) as well as intuition-building components, such as illustrative examples, metaphors, pictures and thought experiments (short: examples). The language model goes over the full-texts block-by-block, and annotates the sample.

After extraction, we have to merge the annotated individual occurrences of examples and positions (which might come from different parts of the same paper, as well as from different papers) into their global types. To do so, I first employ a custom fine-tuned modern-BERT-style model together with an approximate nearest-neighbor (k -NN) search algorithm to determine whether examples are candidates for merging. To decide whether two similar examples/positions should actually be merged, I again employ the LLM. The outcome of this merging process is illustrated in Figure 1, Figure 2 shows the size-distribution of merged clusters.

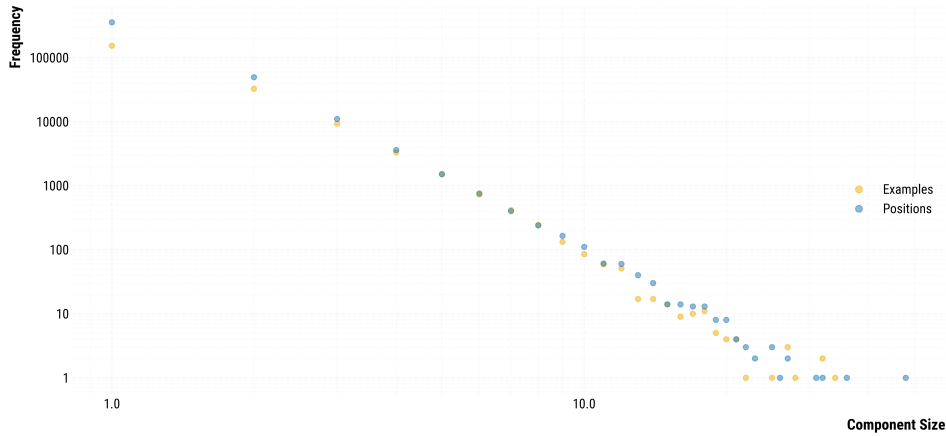


Figure 2: Log-log plot showing the frequency of different component sizes after the merging step, or in other words, how often examples and positions of varying commonalities occur. We note that both examples and positions appear most often only once, without ever being taken up by other philosophers.

After merging positions and examples so that their usage is unified across different papers, we can build a bipartite network, in which examples and positions are linked if they co-occur within the same article (Figure 3). Together with the temporal order of when examples/positions are first introduced into discourse, and potentially data on citations, this dataset provides a unique view into the flow of philosophical reasoning, and a key collective cognitive process, more broadly speaking. I hope that we can answer questions like: How does attention to both philosophical positions and examples wax and wane over time? Are philosophers novelty-driven, preferring to explore new positions and examples, or are they more conservative, focusing on long-lasting canonical cases? But also, do examples emerge before the positions they appear to support, or do positions come first, and enable the discovery of new examples?

At this stage, the data collection and curation for this project is largely con-

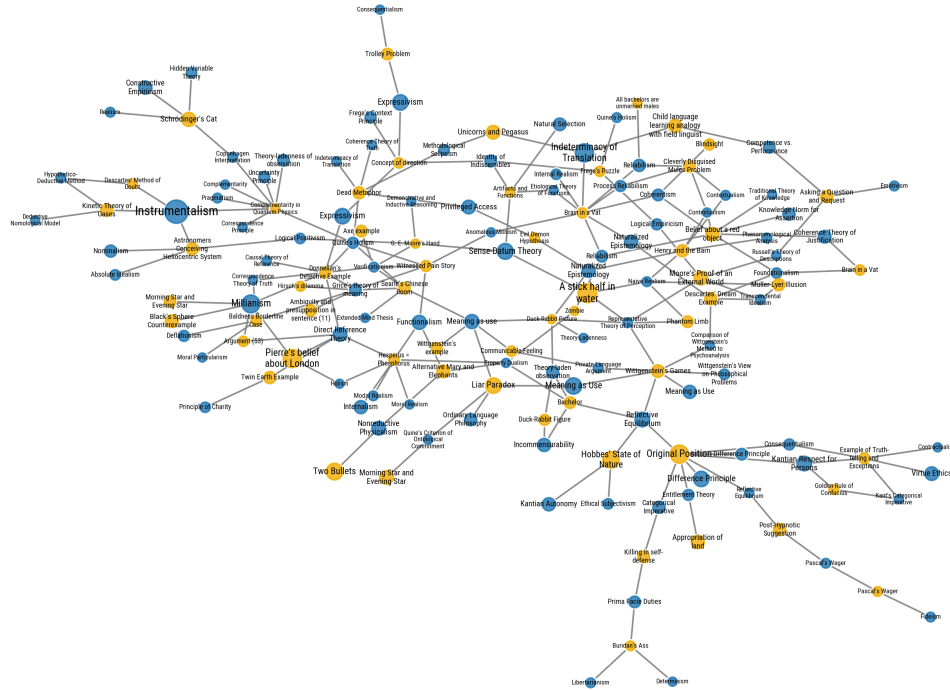


Figure 3: The most common examples (yellow) and positions (blue) in the dataset as a bipartite network. We note that many links are easily interpretable and relate clearly to discussions in philosophy from the second half of the 20th century. This illustration shows only a small excerpt from the actual network, which consists of more than 60,000 nodes.

ducted. The question now is: what are the right models to apply to it to answer the questions above? My hope for taking part in this year's AGWCSS would be to profit from the vibrant community of researchers at the SFI and their expertise in methods of computational social science and network-analysis. This is both with regard to the more technical aspects of this work – e. g., how can we link the somewhat complicated aggregation procedures (k -NN graphs with post-hoc LLM-merging and potentially post-hoc-clustering) back to the classical models from network-theory, and interpret them in light of their properties? What are the best practices, the risks and safeguards in using LLM's the way I do in this study? But the SFI will also be the right place to meet thinkers with whom to discuss the broader strokes of this project, the interplay of formal logic and imagination, and the dynamics of collective intellectual pursuits.