Carnap’s Robot Redux: LLMs, Intensional Semantics, and the Implementation Problem in Conceptual Engineering

Bradley P. Allen

In his 1955 essay ”Meaning and synonymy in natural languages” (Carnap, 1955), Rudolf Carnap presents a thought experiment wherein an investigator provides a hypothetical robot with a definition of a concept together with a description of an individual, and then asks the robot if the individual is in the extension of the concept. In this work, we show how to realize Carnap’s Robot through knowledge probing (Youssef et al., 2023) of an large language model (LLM).

We generalize the approach taken in Allen (2023) to implement the classification procedures proposed by Nado (2023) as targets for conceptual engineering. We define an intensional semantics (Von Fintel & Heim, 2021) for a first-order language without function symbols, where W and D are non-empty sets of possible worlds and individuals, respectively. The intension of a k-ary predicate symbol P is a function from W to the powerset of k-tuples of elements of D. The experimental framework envisioned by Carnap can be implemented using prompt engineering (Liu et al., 2023) of an LLM to define such an intension function for a given concept predicate using a natural language definition of the concept, and then putting the LLM in the role of Carnap’s Robot by applying that function to a natural language description of an individual, yielding a statement indicating if the individual is in the extension of the concept.

This method depends on our ability to trust that the LLM effectively captures the meaning of a given concept (Heersmink et al., n.d.). The question of whether LLMs capture meaning is widely debated (Bender et al., 2021; Kambhampati, 2024; Lederman & Mahowald, 2024). Mandelkern and Linzen (2024) argue that LLMs are indirectly verbally grounded in the language present in their training corpora, and thus capable of a limited form of meaning. Assuming this, we argue that the above method can provide a useful cognitive tool (Menary & Gillett, 2022; Novaes, 2012) for conceptual engineers to compare the extension of a proposed concept definition to the extensional knowledge represented as facts in a given knowledge base (Allen & Groth, 2024). This provides an approach to calibrate trust in an LLM

Bradley P. Allen
University of Amsterdam, Amsterdam, The Netherlands, e-mail: b.p.allen@uva.nl

used in this manner, and can also be viewed as an instance of a corpus method for experimental philosophy (x-phi) (Fischer & Sytsma, 2022; Sytsma, 2023), with relevance to the relationship between x-phi, conceptual engineering, and Carnapian explication (Koch, 2019; Pinder, 2017; Shepherd & Justus, 2015).

We close by arguing that the above method provides a possible solution to the implementation problem in conceptual engineering, which poses the question of whether (re)engineered concepts can be effectively adopted by a population of human speakers (Cappelen, 2018; Jorem, 2021). Online knowledge bases such as Wikidata (Vrandečić & Krötzsch, 2014) have a direct and material impact on society by virtue of their use in online search, discovery, and recommendation (Peng et al., 2023). Using the above method to guide changes to facts in a knowledge base to better align with the extension of a proposed definition provides an indirect method for shifting the semantic meaning of a concept for the specific linguistic subgroup (Matsui, 2024) constituted by users of such online knowledge bases.

References

Allen, B. P. (2023). Conceptual engineering using large language models. *arXiv preprint arXiv:2312.03749*.

Allen, B. P., & Groth, P. T. (2024). Evaluating class membership relations in knowledge graphs using large language models [To appear.]. *European Semantic Web Conference*.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Cappelen, H. (2018). *Fixing language: An essay on conceptual engineering*. Oxford University Press.

Carnap, R. (1955). Meaning and synonymy in natural languages. *Philosophical studies*, *6*, 33–47.

Fischer, E., & Sytsma, J. (2022). Projects and methods of experimental philosophy. *The Compact Compendium of Experimental Philosophy*, 39.

Heersmink, R., de Rooij, B., Vázquez, M. J. C., & Colombo, M. (n.d.). A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness.

Jorem, S. (2021). Conceptual engineering and the implementation problem. *Inquiry*, *64*(1-2), 186–211.

Kambhampati, S. (2024). Can large language models reason and plan? *Annals of the New York Academy of Sciences*.

Koch, S. (2019). Carnapian explications, experimental philosophy, and fruitful concepts. *Inquiry*, *62*(6), 700–717.

Lederman, H., & Mahowald, K. (2024). Are language models more like libraries or like librarians? bibliotechnism, the novel reference problem, and the attitudes of llms. *arXiv preprint arXiv:2401.04854*.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35.

Mandelkern, M., & Linzen, T. (2024). Do language models' words refer? *arXiv preprint arXiv:2308.05576*.

Matsui, T. (2024). Local conceptual engineering in a linguistic subgroup and the implementation problem [Preprint at https://philpapers.org/rec/MATLCE-8].

Menary, R., & Gillett, A. (2022). The tools of enculturation. *Topics in Cognitive Science*, *14*(2), 363–387.

Nado, J. (2023). Classification procedures as the targets of conceptual engineering. *Philosophy and Phenomenological Research*, *106*(1), 136–156. https://doi.org/https://doi.org/10.1111/phpr.12843

Novaes, C. D. (2012). *Formal languages in logic: A philosophical and cognitive analysis*. Cambridge University Press.

Peng, C., Xia, F., Naseriparsa, M., & Osborne, F. (2023). Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, 1–32.

Pinder, M. (2017). Does experimental philosophy have a role to play in carnapian explication? *Ratio*, *30*(4), 443–461.

Shepherd, J., & Justus, J. (2015). X-phi and carnapian explication. *Erkenntnis*, *80*, 381–402.

Sytsma, J. (2023). Ordinary meaning and consilience of evidence. *Advances in experimental philosophy of law*, 171.

Von Fintel, K., & Heim, I. (2021). Intensional semantics [MIT].

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, *57*(10), 78–85.

Youssef, P., Koraş, O. A., Li, M., Schlötterer, J., & Seifert, C. (2023). Give me the facts! a survey on factual knowledge probing in pre-trained language models. *arXiv preprint arXiv:2310.16570*.