Large language models grasp morality in concept

A social constructionist view of ‘modeling morality’

Abstract. Both large language model (LLM) researchers and end-users are increasingly confronted with a series of moral questions, such as: Can and should LLMs express moral or normative judgements [1]? What about racist, sexist, ableist, etc. sentiments [2, 3]? Can and should LLMs ‘take responsibility’ for what they say [4, 5]? Although LLM researchers have sought a wide range of technical solutions, fully answering these questions requires us first to develop a philosophical account of how LLMs can mean (morally).

Accordingly, we provide a general theory of meaning which extends Saussurean [6] semiotic notions to intentionality-agnostic contexts. In contrast with conventional accounts of meaning reliant upon human-specific mental constructs [7], we expand the set of meaning-agents to include objects such as LLMs and social bodies. We understand meaning as an interplay of signs, concepts, and objects, where signs pick out objects through concepts. Meaning takes shape over time through two intertwined processes of concretization and inscription. Through *concretization*, meaning-agents internalize the concepts of objects as informed by external stimuli. This parallels social constructionist accounts of a historicized object whose determinations unfold over time [8, 9]. Through *inscription*, meaning-agents then ‘write’ the relations between concepts back into the world by acting on the basis of internalized concepts.

This theory of meaning helps explicate the precise nature of LLMs as meaning-agents. We understand concretization and inscription in LLMs as being broadly similar to concretization and inscription in social bodies as meaning-agents. LLMs are trained to reproduce signs and objects in datasets assembled from the social body, and therefore crystallize the disagreement in the social fields of discourse which constitute meanings. Accordingly, we suggest that LLMs model the social development of meanings, occupying a position akin to that of philosophical models.

Therefore, LLMs grasp the constructions of human society — such as morality [10], gender [11], and race [12] — in concept, insofar as they are textually articulated.

One of the most pressing research areas on LLMs is *alignment*: how to align LLM behaviors with certain values [13, 14]. Recent work has suggested that LLMs should reflect value pluralism to avoid imposing homogenous values on heterogeneous populations [15, 16]. This work considers value pluralism *in content*, which requires that meaning-agents reproduce sign-object relationships that reflect value pluralism as a worldview. In contrast, we draw attention to pluralism *in concept*, which requires that meaning-agents reproduce sign-object relationships that reflect the diversity and contradictions which structure *fields of discourse*.

Our work sets forth the suggestion that in dealing with questions of morality in digital contexts, we must go beyond treating moral systems as complete, self-consistent, and total. Instead, we must also recognize the complex conditions under which these systems emerge.

References

[1] Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Liang, J., Dodge, J., ... & Choi, Y. (2021). Can machines learn morality? The Delphi experiment. arXiv preprint arXiv:2110.07574.
Abdulhai, M., Creepy, C., Valter, V., Canny, J., Jacques., Natasha. (2023). Moral Foundations of Large Language Models. AAAI 2023 Workshop on Representation Learning for Reponsible Human-Centric AI.

[2] Omiye, J., Lester, J., Spichak, S., Rotemberg, V., Daneshjou, R. (2023). Large language models propagate race-based medicine. NPJ Digital Medicine. doi: 10.1038/s41746-023-00939-z.

[3] Field, A., Blodgett, S. L., Waseem, Z., & Tsvetkov, Y. (2021). A survey of race, racism, and anti-racism in NLP. arXiv preprint arXiv:2106.11410.

[4] Phute, M., Heiblng, A., Hull, M., Peng, S., ..., Chau, D. (2023). LLM Self Defense: By Self-Examination, LLMs Know They Are Being Tricked. arXiv preprint arXiv:2308.07308.
Lu, Q., Zhu, L., Xu, X., Xing, Z., Harrer, S., & Whittle, J. (2023). Building the Future of Responsible AI: A Pattern-Oriented Reference Architecture for Designing Large Language Model based Agents. arXiv preprint arXiv:2311.13148.

[5] West, P., Lu, X., Dziri, N., Brahman, F., Li, L., Hwang, J. D., ... & Choi, Y. (2023). The Generative AI Paradox:" What It Can Create, It May Not Understand". arXiv preprint arXiv:2311.00059.

[6] De Saussure, Ferdinand ; Bally, Charles ; Sechehaye, Albert ; Riedlinger, Albert & Harris, Roy (1987). Course in General Linguistics. Tijdschrift Voor Filosofie 49 (1):125-127.

[7] Grice, Herbert Paul (1957). Meaning. Philosophical Review 66 (3):377-388.

[8] Searle, John R. (1995). The Construction of Social Reality. Free Press.

[9] Hacking, I. (1995). The looping effects of human kinds. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 351–394). Clarendon Press/Oxford University Press.

[10] Nietzsche, F. W. (2009). On the genealogy of morality: A polemic (M. Clark and A. J. Swensen, Trans.). Hackett Publishing.

[11] Butler, Judith (1989). Gender Trouble: Feminism and the Subversion of Identity. Routledge.

[12] Omi, M., & Winant, H. (1994). Racial formation in the United States: From the 1960s to the 1990s (2nd ed.). Routledge.

[13] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

[14] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073.

[15] Sorensen, T., Jiang, L., Hwang, J., Levine, S., Pyatkin, V., West, P., ... & Choi, Y. (2023). Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. arXiv preprint arXiv:2309.00779.

[16] Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., & Bertulfo, D. J. (2022). The Ghost in the Machine has an American accent: value conflict in GPT-3. arXiv preprint arXiv:2203.07785.