COOPERATIVE AGENCY-CENTERED LLMS

Iyadunni J. Adenuga
Department of Computer Science and Technology
Kean University
Union, NJ, USA
{iadenuga}@kean.edu

ABSTRACT

AI models are not readily accepted and adopted in highly consequential fields like education, which cater to societal ideals. This is because injecting human values such as agency, which is an essential ingredient for learning, is hard. Using knowledge from cooperative AI and Liu et al. (2023) Stable Alignment method, we propose (without evaluation) a method for aligning large language models (LLMs) with the human agency of two different groups: teachers and students. This could ensure that effective learning occurs even with LLMs.

1 INTRODUCTION

AI systems are ubiquitous in today's society, but there is still hesitation about their adoption, especially in sensitive areas like education and law enforcement, since they are usually not aligned with societal ideals. This is because injecting human societal values into AI models is difficult. There is also the compounding issue of the systems existing in common sociotechnical environments with multiple groups, usually with diverging goals, values, and ideals. Even though these groups typically need to cooperate to exist peacefully in these environments, there is still the fear that *upper* class groups (i.e., groups with more power) would be treated more favorably than the *lower* class. For example, consider the school environment. For effective learning to occur, there has to be cooperation between the human teacher, human student, and the models that represent them. Also, each entity requires agency (an innate psychological need) for a fulfilling experience. How can an AI model adaptively allow for the agency of different groups or align with the values of multiple groups while promoting cooperation in a socio-technical environment? AI models that prioritize the collective agency of each group while also ensuring cooperation could be well-accepted in society. This is because the welfare of all groups could be guaranteed.

Using the school environment as a case study and knowledge from Co-operative AI, this paper proposes a method for ensuring AI models align with the human need for agency regardless of the group they belong to.

2 AGENCY AND LEARNING

Agency refers to the human innate trait to control and influence outcomes in a specified environment. It is closely related to the psychological need for autonomy (Bennett et al., 2023). Bennett et al. (2023) conducted an extensive examination of past human-computer interaction (HCI) research works on agency and autonomy, highlighting their essential positive contributions to user experience and well-being, such as in forms of improved "satisfaction", "meaningful communication", "trust", and "learning". Students who feel in control of their learning experience, even in low amounts, have been shown to be more engaged and interested, which results in better performance (Taub et al., 2020). There is a limit to the positive effects of agency as the level increases, necessitating the importance of the presence of a balancing teacher's control (Taub et al., 2020; Rajala et al., 2016a).

3 CO-OPERATIVE AI

Co-operative AI is a branch of AI research that extends the bi-directional human-AI alignment to multiple groups. Important elements that influence co-operation (Dafoe et al., 2020) include "communication", "commitment", and "institutions". A school environment embodies these influences and so can help construct the policies governing the cooperative AI models in that environment. Teachers, students, and administrators usually agree or negotiate a set of norms and are committed to effective learning, but each group still requires agency to function properly.

4 LLM SOLUTION

Co-operative AI and models trained on simulated social interactions have been introduced as methods for creating AI models that center and align with human-centered values. For easier access to experts, the domain for this approach would be the school environment, and we would focus on two main groups: teachers and students. Teachers are usually against the use of LLMs/generative AI because they don't believe students gain the requisite knowledge. Students favor these systems due to their ease of use when seeking assignment help. Teachers often also perform a mentoring role for students. They assess a student's current knowledge state and impart new knowledge based on the assessment.

As described in Section 2, effective learning requires that the LLM serves the student's agency needs (through reflection) while also ensuring teacher control (through level-appropriate response to the student's inquiry). Since we also desire cooperation (See Section 3) among these two groups, there must be prior agreement to a set of norms such as a curriculum, commitment to the goal of ensuring students' learning, and provision of a negotiation space for the teacher-student agency levels. The negotiation space may be facilitated by the LLM so that the teacher and student groups adapt to the model.

On the other alignment end, what if an LLM generated "guiding responses" that help the students actively engage with the concept being learned? Such an LLM would allow for adaptive, "generative", and "transformative" interactions (Rajala et al., 2016b). This approach seeks inspiration from the Liu et al. (2023)'s *Training Socially Aligned Language Models on Simulated Social Interactions* Stable Alignment method. Liu et al.'s method simulates numerous social interactions which form data that are used to train an alignment LLM. The collected interaction data, in addition to aligned and misaligned data, also contains observer ratings, feedback, and iterative responses. The data generation is guided by a *Sandbox Rule*. Similarly, we want to simulate cooperative social interactions that cater to the agency of both teachers and students.

For our school environment case, first, qualitative interviews with teachers and students in different programming class levels are important. These interviews would be used to determine common learning levels and associated characteristics of responses for each level in our *simulated society*. These level-appropriate responses would be used to define the characteristics of a "guiding response" such that it could lead to a high agency score on a 7-point Likert agreement scale. (Teacher: How well are you in control of this learning environment?; Student: Are you in charge of this learning process?) As a *Sandbox Rule*, the best "guiding responses" generate high agency scores for both teacher and student groups. This interview data would further help in the construction of the simulated society with teacher and student social agents with different attributes (i.e. initial memory systems). There would also be teacher and student observer agents that provide reflections and control scores. A social agent LLM would generate a response to a given programming question and would update its answer/response based on the current programming level and agency scores from the teacher and student agents. This would generate a cooperation-based alignment dataset. An item in the cooperation-based alignment interaction data would contain a programming question, the student's thoughts/reflection on the question (which should be translated to a student control score), their programming level, level-appropriate response, and information about progressive steps to the current programming level. An AI model that learns from this alignment dataset will be created. The aim is to optimize for the best "guiding" responses with the possible highest student and teacher agency scores.

5 CONCLUSION

The aim of this paper is to introduce a method based on (Liu et al., 2023) for aligning an AI model to the agency needs of two different groups: teachers and students using principles from cooperative AI such that students' learning is improved. This method requires extensive improvement and evaluation but it introduces questions about the possibility of the "human" in human-AI alignment field containing different human stakeholders with different roles and interests.

REFERENCES

Dan Bennett, Oussama Metatla, Anne Roudaut, and Elisa D Mekler. How does hci understand human agency and autonomy? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2023.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630*, 2020.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. Training socially aligned language models on simulated social interactions. *arXiv preprint arXiv:2305.16960*, 2023.

Antti Rajala, Kristiina Kumpulainen, Anna Pauliina Rainio, Jaakko Hilppö, and Lasse Lipponen. Dealing with the contradiction of agency and control during dialogic teaching. *Learning, Culture and Social Interaction*, 10:17–26, 2016a.

Antti Rajala, Jenny Martin, and Kristiina Kumpulainen. Agency and learning: Researching agency in educational interactions. *Learning, Culture and Social Interaction*, 10:1–3, 2016b.

Michelle Taub, Robert Sawyer, Andy Smith, Jonathan Rowe, Roger Azevedo, and James Lester. The agency effect: The impact of student agency on learning, emotions, and problem-solving behaviors in a game-based learning environment. *Computers & Education*, 147:103781, 2020.