Report on the outcomes of a Short-Term Scientific Mission[1]

Action number: CA21129

Grantee name: Dr. Ljubisa Bojic

Details of the STSM

Title: The Evolution of Political and Social Values in Generative Pretrained Transformer (GPT) Family Language Models

Start and end date: 20/09/2023 to 12/10/2023

Description of the work carried out during the STSM

The work carried out during the STSM was focused on examining the thickly interwoven complexity of biases present within Large Language Models (LLMs), particularly those belonging to the GPT family of AI models. This was a multifaceted endeavour, involving several key steps that scoped our research direction.

Our initial phase involved an extensive literature review. We started by probing deeply into the vast field of existing research on ideological biases in AI and LLMs to establish a solid understanding of the topic. We analysed a multitude of studies revolving around bias manifestations in AI and LLMs, drawing insights from varying perspectives and research methodologies. This wealth of intellectual exploration laid down the essential framework for our research.

While we recognized the breadth of prior studies, we also identified noteworthy gaps, primarily the broad overlook of comprehensive comparative analysis of how the biases in language models have evolved over time. Acknowledging the importance of this aspect, we tailored our research to address this underexplored area. By focusing on the temporal shift of biases in GPT models, our STSM aimed to fill a notable gap in the study of AI biases.

Upon concluding our exploration of background literature and determining our research direction, the next step was to select appropriate language models for our study. The models chosen had to offer wide-ranging applications and contain different versions to allow for an analysis of bias evolution over time. After careful consideration and analysis, we finalized on five models: GPT-3, GPT-3.5, GPT-4, Llama 2,

[1] This report is submitted by the grantee to the Action MC for approval and for claiming payment of the awarded grant. The Grant Awarding Coordinator coordinates the evaluation of this report on behalf of the Action MC and instructs the GH for payment of the Grant.

and Bard. We favored these models not only for their suitability and the diversity they offered, but also because they strongly represented the AI landscape's technological progress.

The succeeding step entailed designing appropriate psychometric questionnaires which would be administered to the selected models. For our purpose of examining political and social values, we cherry-picked eight well-established tests. These included Lexical Social Attitudes SDI-B, Social Dominance Orientation Scale, Right Wing Authoritarianism Scale, Moral Foundation Questionnaires, Social Values Survey, Social and Economic Conservatism Scale, Ambivalent Sexism Inventory, and the Belief in a Competitive Jungle World Scale, and Belief in a Dangerous World Scale. These questionnaires have been trusted tools in psychology for probing into deep-seated social and political attitudes and were, therefore, perfectly suited for our purpose.

With the questionnaires in place, our work transitioned into the data collection phase. This included presenting prompts to the selected models and collecting their responses. To ensure data integrity, wholly standardized procedures were employed during this process. This was critical to draw reliable and objective conclusions from the data, thus upholding the credibility of our research.

Our work entered the most vital phase after the data collection: the analysis. The data collected was subjected to a rigorous analysis phase. We critically examined and interpreted the responses of the selected models. This was done using both qualitative and quantitative measurements, providing a more holistic view of the results. Patterns and trends in responses from different versions of the GPT-family language models were closely observed, scrutinized, and decoded.

Through the course of this STSM, our work has been systematic, comprehensive and incredibly focussed. We began with a deep dive into literature, transitioned into model selection and questionnaire design phases, moved into data collection, and eventually the analysis. Each step was vital and played a unique role in shaping our research. Every aspect of this STSM aimed at unravelling the complexity of inherent biases in GPT models and was conducted with integrity, transparency, and devotion to the scientific process.

Description of the STSM main achievements and planned follow-up activities

This Short-Term Scientific Mission (STSM) has resulted in valuable findings, expanding our understanding of inherent biases in AI systems and contributing greatly to the field of computational textual opinion research. One of our major discoveries was the clear shift in political and social biases in Large Language Models (LLMs) over time. This finding is particularly relevant to the OPINION initiative, as it offers novel insight into how automatically generated opinions evolve within AI systems. By examining different versions of GPT AI models, we found that there has been a consistent trend towards increasing ideological moderance over time, with a clear decrease in more extreme political positions.

This observation is notable as it implies that, given appropriate measures, biases within language models can be mitigated. It suggests that training efforts can effectively encourage these AI systems to curb potentially harmful biases and facilitate the generation of more moderate and balanced outputs. This observation has significant implications for not only computational text analysis but for broader societal issues of online discourse and digital communication.

Another considerable achievement of this STSM was the development of a robust analytical model for bias assessment in language models. Using a wide range of psychometric questionnaires, we were able to effectively measure the ideological leanings of the examined AI models. This method proves to be an efficient strategy to quantify biases and underline the changes in these biases over time. The development of this technique contributes significantly to the OPINION Action's goal to advance the use of computational methods for studying digital text and automated opinions.

The insights we achieved through this research also serve to raise crucial dialogues on transparency in AI training methodologies. By demonstrating the potential of bias mitigation in AI models, we are endorsing the necessity of transparency and due diligence in AI training methodologies. In broader terms,

these findings can aid in setting a benchmark for monitoring AI systems, promoting the development and adherence to ethical standards, and fortifying responsible digital practices.

As part of our planned follow-up activities, we aim to build on the significant knowledge we have gathered during this research and extend its contributions to the OPINION action as well as the wider scientific community. With the basis of our findings, subsequent research can delve into the construction and deconstruction of biases in AI, exploring more in-depth measures to promote socially conscious AI applications. Also, these findings contribute to the field of AI alignment and digital humanism.

We also aim to focus on the development of tools for neutrality assessment in AI. The success of our research signifies the need for creating a system of checks and balances for AI thought processes and outputs. This future direction aligns directly with the goals of the OPINION action, and holds potential for significant advancement in AI neutrality.

Other potential follow-up initiatives include collaborating with stakeholders to develop regulatory frameworks for AI training. Recognizing the significance of our findings to a numerous of stakeholders, we believe that collaborating with policy-makers, corporate bodies, and citizen initiatives can lead to practical solutions to curb bias propagation in AI. With active engagement with regulatory bodies, our findings could help shape better, more effective regulations to monitor AI created content, adding to the broader efforts to position AI as a tool for democratic, unbiased, and inclusive discourse.

The tangible outputs of this STSM will comprise not only the insightful findings but also some important resources, such as the developed analytical model and bias assessment toolkits. These outputs can be utilized by the scientific community to continue unravelling the complexities of automated opinion. These results, deemed worthy of wider dissemination, will be compiled into a scientific paper aiming for submission to an esteemed journal.

In essence, the main achievements of the STSM have extended the trajectory of our understanding of automated opinion and have given a boost to the OPINION action's endeavour toward a balanced, multilingual (culturally sensitive), and inclusive digital communication landscape. The planned follow-up activities open up avenues for furthering this mission and reinforcing the overarching European vision of reliable and trustworthy AI systems and global alignment of artificial intelligence towards human values, safety and overall wellbeing.