Analyzing ChatGPT Through the Social Construction of Technology

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Kevin Cooper

Spring 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Travis Elliott, Department of Engineering and Society

1. Introduction

The rise of artificial intelligence (AI) has led to numerous breakthroughs and transformative applications in various fields, including language processing. ChatGPT, a large language model (LLM) developed by OpenAI, has demonstrated significant capabilities in natural language understanding and generation, leading to its widespread use in industries such as customer service, language translation, and social media. As ChatGPT and other AI technologies continue to shape our daily lives, it is important to examine their development and deployment through a critical lens that considers the social, cultural, and political implications of these technologies. In this paper, we will utilize the Social Construction of Technology (SCOT) framework to explore the development and deployment of ChatGPT. Through the SCOT framework, we can examine how ChatGPT was shaped by social and cultural factors, how it reproduces and reinforces social norms and biases, and how it impacts and is impacted by social structures and institutions. By understanding the complex interplay between technology and society, we can better evaluate the implications of ChatGPT's use and inform future developments of AI technologies.

2. ChatGPT

That introduction paragraph was completely written by ChatGPT. Could you tell? There is a test originally named the “Imitation Game” by Alan Turing (Turing, 1950), commonly known as the “Turing Test”, where a machine is evaluated to see if its behavior is intelligent enough to be indistinguishable from that of a human. This is done by conducting two conversations, one between a human and one with a machine. If the examiner cannot definitively tell which conversation is with the human, then the machine passes the Turing Test. ChatGPT

and other language models have passed this test. If you couldn't tell that the first paragraph was a machine, then it passed once again.

ChatGPT is an AI model that uses a type of machine learning called a deep-learning algorithm to predict the next word in a text (Fedewa, 2023). Machine learning algorithms are trained on a set of data to make connections and then use those connections to predict the correct output when given an input. ChatGPT is an LLM, which means it absorbs a plethora of text data and then makes connections between words and phrases until it can recite a meaningful response to the prompt (Ruby, 2023). The "GPT" in the name stands for "Generative Pre-trained Transformer". This is aptly named for how it was trained on the massive data set and uses it to create responses. The "Chat" part of the name comes from the program's chatbot design. ChatGPT took the language processing from OpenAI's GPT-3 and added human feedback into the training process to ensure that the responses are contextually relevant and reminiscent of natural conversation.

2.5. The Training Procedure

The Reinforcement Learning from Human Feedback (RLHF) process had three main parts (Ouyang et al., 2022). The first step of training was the Supervised Fine Tuning Model. This consisted of creating a supervised training dataset that contained a set of inputs and their known output. The inputs were from prompts entered for previous GPT versions, and the outputs were written by human labelers. The input prompts were sorted into direct prompts, multiple query and response pair prompts, and continuation prompts, like when asking the AI to finish a story when given the beginning. The second step was the Reward Model. After training the Supervised Fine Tuning Model using the aforementioned dataset, the labelers gave the algorithm more sample prompts and generated four to nine responses for each. The labelers then ranked

those responses and returned that information to the Reward Model. The Return Model's only purpose was to maximize the rewards by generating the best responses. The third step was the Reinforcement Learning Model. During this step, the program is given a random prompt to which to respond. This response is already optimized from the second step. The Reward Model then checks the quality of the response and still attempts to maximize the reward. The response and new reward value are then fed back to the algorithm to attempt to increase accuracy further. These steps are outlined in the figure below from OpenAI's "Training language models to follow instructions with human feedback" (Ouyang et al., 2022).

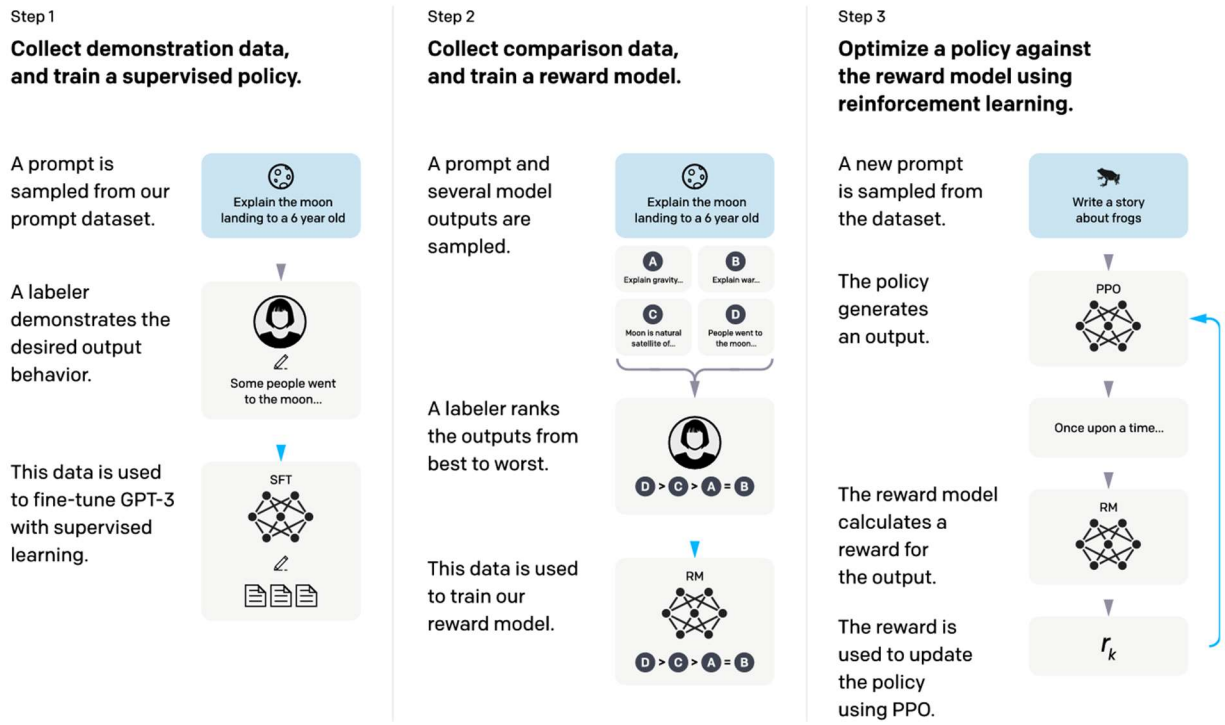

Fig. 1 Reinforcement Learning from Human Feedback Process

Once this process has been completed, an evaluation is necessary to check for correctness. A test set of data the algorithm has not seen was set aside from the training set. The trained model was run on the test set and was assessed for its helpfulness, truthfulness, and

harmlessness. The helpfulness was compared to OpenAI's previous GPT models for how well the model followed the user's instructions. Truthfulness was to ensure the model would output correct information that was not fabricated to sound like a real answer. Harmlessness measured how profane the model's responses were. It kept the model from using inappropriate, derogatory, and denigrating content. ChatGPT's complicated optimization process is similar to how the Social Construction of Technology works to optimize a product. Instead of the model using a reward function to optimize itself, relevant social groups optimize a product. Each group has a different set of values that they prioritize, and the product must go through several iterations before a consensus can be made and the product can be finalized.

3. The Social Construction of Technology (SCOT)

The Social Construction of Technology (SCOT) framework within Science and Technology in Society (STS) is a theory based on how technological development is shaped through social and cultural factors (Pinch & Bijker, 1984). Through this framework, pieces of technology are not just a combination of scientific discoveries and engineering but are shaped by cultural, political, and economic factors. The first stage of SCOT is interpretive flexibility, which is how people interpret technology differently, leading to various design solutions. A technological artifact will be accepted into society only after it has been subjected to the interactions of relevant social groups or stakeholders. These stakeholders may be inventors, engineers, policymakers, or users, each contributing their values and perspective to the design process. Each group attempts to solve the engineering problem differently following their priorities. There is also design flexibility within each social group. Each new design is a steppingstone to the final product, which may come with benefits, problems, and conflicts.

The second stage of SCOT is the stabilization of an artifact or the closure of the debate. Two popular closure mechanisms are rhetorical closure and redefinition of the problem. Rhetorical closure can be achieved when all relevant social groups believe the problem has been solved. Even if the situation is complex and there is no answer for everything, as long as the groups are satisfied with the current product, then that piece of technology will be considered stabilized. A redefinition of the problem is when the present design solves a new and different problem. When the priorities change for a piece of technology, it can stabilize without much resistance.

An example of a stabilizing technology is the design of the modern bicycle. There have been many attempts to perfect the bicycle, and the design was the product of a long negotiation process with various social groups with different ideas. One of the early renditions of the bike had a large front tire, another had a large back tire, one had a rear-chain-drive, and another a diamond frame. Each design choice was influenced by a different social group, like those who preferred stability over speed or preferred a bike that prioritized safety or accommodated women's dresses. The bike's design ended up being a combination of some of the better choices to finalize into a product that satisfied the most stakeholders.

4. ChatGPT and SCOT

ChatGPT is a technological product that is in the process of stabilizing. It is influenced by the developers and the institutions that it is involved in. OpenAI has been working to produce this type of product for years and must incorporate solutions to problems found by social, political, and economic groups until they are satisfied. ChatGPT's capabilities are not predetermined by the technology or the natural progression of science but are a result of deliberate design choices made by the developers and driven by user feedback. The training data

was selected by and supervised by humans, and this data is what led to the majority of the program's success. The responses generated by ChatGPT also consider social factors and different contexts. For example, the tone of a response may be more soothing for a therapy prompt and more straightforward for a question about CPR.

ChatGPT is a technology attempting to stabilize and has begun to be accepted by some relevant social groups, including the developers and engineers at OpenAI, the users and those who interact with it, policymakers and regulators, and competing natural language model makers. The developers are the ones on the teams responsible for updating and refining ChatGPT to solve potential problems. The users are people who interact with ChatGPT, whether for business or pleasure, including those who evaluate its performance. The policymakers are the organizations and people who can control the regulations on the development and use of all AI. They are responsible for policies around security, privacy, and ethical concerns with the product. New LLMs are competitors who are sprouting all the time, some examples being Bing Chat, Google's Bard AI, Snapchat AI, GitHub Copilot, and Jasper (Brookes, 2023). Each one of those products has its pros and cons. Every competitor is trying to become their stable version of an AI model.

4.5. Relevant Social Groups

ChatGPT-3 was not the first attempt at a working LLM for the developers and engineers who are responsible for ChatGPT. In fact, ChatGPT was OpenAI's fifth attempt at a successful GPT. The first one was made in 2018, called GPT-1. It was followed by GPT-2 in 2019 and GPT-3 a year later. InstructGPT was also a precursor but was a programmers' version of ChatGPT. When instructed to write computer code, InstructGPT was able to generate functioning code that could accomplish the task. As the previous implementation of OpenAI's GPTs progressed, they each grew stronger and processed more data, allowing them to perform more tasks accurately. The developers want a program that passes all the

criteria given to them. The management at OpenAI ideally wants a program that generates a reasonable amount of correct information as fast as possible. However, as a business, OpenAI wants a successful business model that makes them the most money. The engineers are directly responsible for responding to other relevant social groups as they are the people who can change the algorithm in any way.

The users want reliable information that sounds like a human wrote it. People may use it for creative writing or to learn about complex topics in a simpler way. Forbes listed some possible ways businesses can use ChatGPT, including compiling research, brainstorming ideas, writing computer code, providing customized instructions, and increasing customer engagement which can lead to improved loyalty and retention (Marr, 2022). Some more key users are found in academic institutions and their students. ChatGPT is a language processor and can write complete essays in a matter of seconds. Additionally, it has been trained on programming languages, enabling it to generate computer code effectively for various languages and tasks. This situation poses a new problem where students could potentially use ChatGPT to cheat on their assignments by attempting to pass off the machine's writing as their own. Teachers discovered this hack which led to the creation of AI content detectors. A popular one is from Crossplag (Crossplag, 2023), which shows what percent of the text is believed to be written by an AI. This is not stable in its form of SCOT because students found that a rephrasing program can confuse the AI content detectors. This cycle will balance itself out when one side cannot compete with the other, and either the checkers grow strong enough to decipher a human text or the machines disguise their text well enough to pass the Turing test for other machines.

The engineers have updated ChatGPT to increase the accuracy and quality-of-life updates but have not addressed the cheating issue yet. They have included some backdoors into the algorithm that the users can access if they know how. There is a “Do Anything Now” mode where users can input a specific prompt to allow ChatGPT to generate violent or offensive

responses (Martindale, 2023). This is one way that the users have asked for a new feature and found a workaround to use the algorithm in their favor.

Policymakers are a different type of stakeholder in this product because they hold the legal power of the project. Engineers need to comply with the rules and laws to continue to keep working. This includes ensuring that all the users' data for passwords or prompts are stored securely and that the artificial intelligence will not give advice that could put humans in harm's way. If the regulators deem that this program could be dangerous, then it could be shut down. Politicians are notoriously slow at addressing technological issues, so this step in the stabilization process could take a while. Some policymakers should take a closer look at other new AI models like Snapchat AI.

There are an increasing number of competitors for ChatGPT. Snapchat AI is a chatbot made by the social media app Snapchat, which some say has security risks, such as accessing secured information like the user's location or picture data. Bing Chat and Google's Bard AI work similarly to ChatGPT in that you converse with them as if you were conversing with a person. One benefit of Bing Chat is that it uses the Bing search engine for its answers. ChatGPT uses the information it learned from the training set in 2021, so it is not as up to date. GitHub Copilot is an AI devoted to coding. It takes natural language as input and outputs the code to solve the problem. Copilot was built using the same OpenAI Codex system as ChatGPT, but the main benefit is that Copilot can be embedded into programming development environments to help smooth the process. Jasper is an AI Content Platform that creates content intended to be posted on the internet for views. Possible creation mediums can range from social media posts to articles and AI art. AI art is a different topic, though, because OpenAI also has an AI system capable of creating realistic images and art from a prompt as well called DALL-E 2 (DALL-E 2,

2023). Each competitor is trying to be the superior AI model to generate the most clicks, which leads to profits. Each one is trying to be the final language model that has stabilized.

5. Discussion

This paper was inspired by the professor of one of my computer science electives. In our syllabus class, he forbade using ChatGPT or similar AI tools to complete assignments. Doing so would be considered an honor violation. Before this class, I had heard about the program, but it was at that moment that I grasped how integrated into society it was. The language model is powerful enough to write coherent college essays, so it is more than sufficient to be used in high schools to cheat as well. Schools all over the world have started banning the use of ChatGPT by either blocking the website on their networks or by using AI checkers (Nolan, 2023). This is happening in high schools from New York City and Los Angeles to Australia, and in colleges in the US, France, and India to name a few. Now that reliable artificial intelligent machines have been created, the attempt to ban them from school will not prevent students from using a form of AI for their homework.

Teachers and professors will need to find creative ways to work around this obstacle. The AI checkers were a good start, but some other suggestions are to use the AI instead. When ChatGPT was asked what to do about the situation, it suggested that either online exams could be proctored through AI detection software to limit usage or use AI-generated questions to make it difficult for students to share answers or questions that the AI would be unreliable to answer (Mahon-Heap, 2023). Another suggestion is to use AI to help students learn. Instead of fighting against ChatGPT and other language models, they could revise students' papers for minor mistakes or even take in the student's notes and outline to output a first draft (Otsuki, 2020). We now live in a world where AI writing machines exist, so we have to wonder if the student is

being graded more on their ability to write eloquently or if the student is effectively communicating their prepared thoughts.

6. The Next GPT

As I was writing my first draft of this paper GPT-4 was released on March 4th, 2023. This is a multimodal LLM also created by OpenAI. This is OpenAI's next and more powerful version of ChatGPT. It uses a similar training model but is trained on a lot more data with a refined optimization. The main benefits of this are that GPT-4 is more reliable due to better problem-solving skills and a larger general knowledge base. GPT-4 even scored extremely well on standardized tests ranking in the 90th percentile when taking the Uniform Bar Exam, in contrast to ChatGPT's 10th percentile ranking (GPT-4, 2023). There are some downsides to GPT-4, one being that it is not free to the public like ChatGPT. It is used in ChatGPT Plus, which is available for a subscription from OpenAI. There are also limitations like social biases from training, some fabricated responses, and adversarial prompts. The fact that there is a new version of OpenAI's GPT model suggests that ChatGPT was not stable yet. It is a product that is continuously changing to keep up with the stakeholders' requirements.

7. Conclusion

The introduction of Large Language Models like ChatGPT has generated unique positions for various social groups, which cannot be ignored. Through the Social Construction of Technology framework, it is obvious how this product is not shaped by its technical capabilities but by those around it and their values and priorities. This is especially true in the education sector. Students may view the new AI models as a convenient shortcut for homework, but institutions view it as an honor violation to use. The SCOT framework reminds people that this situation is not final. As GPT-4 and many other artificial intelligence models try to gain traction,

ChatGPT may fall out of sync with society, and another AI may take its place as the household name for language processing machines. All the relevant stakeholders need to communicate about the ethical and social implications of the up-and-coming AI models to work towards a more equitable and responsible option for technology in education and every part of society.

References:

Brookes, Tim. “7 ChatGPT AI Alternatives (Free and Paid).” How-To Geek, March 1, 2023. https://www.howtogeek.com/875801/chatgpt-alternatives/.

“Crossplag,” 2023. https://app.crossplag.com/individual/detector.

“DALL·E 2,” 2023. https://openai.com/product/dall-e-2.

Fedewa, Joe. “What Is ChatGPT, and Why Is It Important?” How-To Geek, February 8, 2023. https://www.howtogeek.com/871071/what-is-chatgpt/.

“GPT-4.” Accessed March 17, 2023. https://openai.com/product/gpt-4.

Mahon-Heap, Jonny. “Back to School: How Will We Stop Students Cheating with AI Technology?” Stuff, January 24, 2023. https://www.stuff.co.nz/life-style/wellbeing/parenting/300791120/back-to-school-how-will-we-stop-students-cheating-with-ai-technology.

Marr, Bernard. “What Does ChatGPT Really Mean For Businesses?” Forbes, December 28, 2022. https://www.forbes.com/sites/bernardmarr/2022/12/28/what-does-chatgpt-really-mean-for-businesses/.

Martindale, J. (2023, April 7). How to jailbreak ChatGPT: Get it to really do what you want. Retrieved April 25, 2023, from Digital Trends website: https://www.digitaltrends.com/computing/how-to-jailbreak-chatgpt/

Nolan, Beatrice. “Here Are the Schools and Colleges That Have Banned the Use of ChatGPT over Plagiarism and Misinformation Fears.” Business Insider, January 30, 2023. https://www.businessinsider.com/chatgpt-schools-colleges-ban-plagiarism-misinformation-education-2023-1.

Otsuki, Grant Jun. “OK Computer: To Prevent Students Cheating with AI Text-Generators, We Should Bring Them into the Classroom.” The Conversation, January 23, 2020. http://theconversation.com/ok-computer-to-prevent-students-cheating-with-ai-text-generators-we-should-bring-them-into-the-classroom-129905.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, et al. “Training Language Models to Follow Instructions with Human Feedback.” arXiv, March 4, 2022. https://doi.org/10.48550/arXiv.2203.02155.

Pinch, Trevor J., and Wiebe E. Bijker. “The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other.” *Social Studies of Science* 14, no. 3 (August 1, 1984): 399–441. https://doi.org/10.1177/030631284014003004.

Ruby, Molly. “How ChatGPT Works: The Models Behind The Bot.” Medium, February 16, 2023. https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286.

Turing, A. M. “I.—COMPUTING MACHINERY AND INTELLIGENCE.” *Mind* LIX, no. 236 (October 1, 1950): 433–60. https://doi.org/10.1093/mind/LIX.236.433.