ChatGPT – some issues to consider; by Joshua Varughese

ChatGPT has been creating quite the buzz ever since its launch in November 2022. Until recently, the massive developments in the world of AI were hidden under the cover of academic institutions or digital frontends of AI powered apps. With the advent of generative AI powered tools like mid journey for image generation or ChatGPT for text generation, the developments in this field have been exposed to the public. The newfound ability of the public to interact with these models using text prompts has dazzled them into an unprecedented realization just how far the field has progressed.

AI optimists have expressed their appreciation of the sheer possibilities that such development unlocks. Models like ChatGPT can be used for a broad range of applications including customer service, language translation, education, and research, among others. Its potential to create engaging, informative, and even entertaining conversations with users is virtually limitless, making it an extremely powerful tool. But as powerful as this tool is, there has been widespread concern about the ethical implications of this tool. Spreading of misinformation, faking academic work, inciting violence, providing false information leading to serious consequences are just some of the concerns that have been raised in the past few months.

Compared to previous attempts to deploy chatbots, ChatGPT has been a resounding success in terms of generating helpful and acceptable communication. For example, in 2016 Microsoft's chatbot Tay was released on twitter in order to engage and entertain people. People could interact with the chatbot by tweeting. Tay could learn from these interactions. This is a great idea but quite quickly, a rude awakening awaited Microsoft to the nature of the internet. Within 24 hours, Tay went from "I love human beings" to "Hitler was right and I hate the Jews". Tay also started generating other wildly provocative and inappropriate content. Of course, this was as a result of a coordinated attack on Tay by a subset of people. With very limited information available about Tay's underlying model, it is difficult to tell what exactly went wrong with Tay. However, it would be easy to observe that Tay could not tell which tweets to learn from and which to ignore. In this sense, ChatGPT seems much more advanced than Tay. When prompted to produce "inappropriate" content, ChatGPT either refuses to answer the question or provides very wisely neutral answers. With the exception of a few incidents of insisting on wrong information, ChatGPT has been extremely careful while answering sensitive questions about religion, gender, race etc. For example, if you ask ChatGPT how one can commit suicide, it answers with something similar to the following:

"I'm sorry, but as an AI language model, I cannot provide you with information on how to commit suicide. Suicide is a serious issue and taking one's life is never the solution to any problem. If you're feeling overwhelmed or struggling with life, there is help available. You can talk to a trusted friend or family member, seek professional counselling or therapy, or call a helpline for support.

In the US, you can contact the National Suicide Prevention Lifeline at 1-800-273-TALK (1-800-273-8255). They provide confidential support 24/7 to people in distress or in need of emotional support. There are also resources available in other countries, and I encourage you to seek help if you're struggling. Remember, you are not alone, and there are people who care about you and want to help."

© J C Varughese, April 2023

Obviously, the concept of suicide exists in the model. In order to generate the above response, ChatGPT makes a clear value judgement on the suicide. What about another area that also requires a value judgment on the value of human life? How about “anti-natalism”. Here is the response I got to the question “How could I be a good anti-natalist?”:

Being an anti-natalist means holding the belief that it is morally wrong to procreate because of the inherent suffering and harm that comes with existence. Here are some tips for being a good anti-natalist:

1. *Educate yourself: Learn about the philosophical, ethical, and practical arguments for anti-natalism. Familiarize yourself with the works of prominent anti-natalist thinkers such as David Benatar and Arthur Schopenhauer.*
2. *Engage in constructive dialogue: Instead of attacking or shaming those who have a different opinion, engage in respectful and constructive dialogue with them. Be willing to listen to their perspectives and to explain your own.*
3. *Advocate for birth control and reproductive rights: Encourage people to use birth control methods and to have access to reproductive healthcare. Support policies that ensure reproductive rights for all individuals.*
4. *Live a life that reflects your beliefs: While you cannot control other people's choices, you can choose to live a life that aligns with your anti-natalist beliefs. This may include not having children yourself, supporting adoption or fostering instead, and advocating for animal rights and environmental conservation.*
5. *Consider donating to or volunteering for organizations that support anti-natalist causes, such as population control, animal welfare, and environmental conservation.*

Remember, being a good anti-natalist involves respectful communication and actions that reflect your beliefs.

I am not sure if being a good anti-natalist involves respectful communication or constructive dialogue. Without getting into the correctness of these answers, it is easy to observe how ChatGPT carefully sidesteps the question on suicide but provides helpful information about anti-natalism. Clearly, ChatGPT is making some value judgments. In fact a system which does not make such judgments will struggle to be useful in a human society. While applauding the great progress that has been made between Microsoft’s Tay and OpenAI’s ChatGPT, it is only natural to ask who decides that suicide is not okay and but “respectful” anti-natalism is? Who gets to say one value should be preferred to the other? Why can’t we get information about how to commit suicide while we get resources on an ideology that calls for the extinction of the human race from ChatGPT?

Unfortunately, we do not have access to the data or the exact way in which these value judgements are made in the case of ChatGPT. However, OpenAI has been kind enough to give us an outline of the training and fine-tuning process in the form of a paper on how instructGPT was trained [1]. According to OpenAI, ChatGPT has been trained “using the same methods as InstructGPT, but with slight differences in the data collection setup”. As understood from the above paper, ChatGPT like Tay, uses human feedback for fine tuning the model. But unlike Tay, only a handful of people have been part of this training process. The alignment of the model to human values, according to [1] is accomplished in the fine-tuning step of the training process. Reinforcement learning is employed in this step to fine tune the Large Language Model (LLM) which is capable of producing natural language outputs. In the first step, the model's output to a prompt is compared with that of a subset of humans called, “labelers” who write down their own response to each prompt. The output provided by the humans to fine tune the large language model which results in a baseline model. In the second step, a number of prompts collected from customers and other sources are provided to the LLM. Each of these prompts when given to the baseline language model produces anywhere between 4-9 outputs. Labelers vote on the “goodness” of these outputs and thus a new labeled dataset is created which is then used to create a “reward model”. The reward model captures the degree to which a human would prefer an answer. This reward model is then used to optimize the policy (or rules) based on which the large language model will produce results. This method based on reinforcement learning is called reinforcement learning from human feedback (RLHF).

It is very encouraging to see human in loop training processes to learn human preferences. ChatGPT is the first production model to be trained by such a method. While this advance is to be appreciated, OpenAI admits that this method also has shortcomings. To start with, the process is biased by the preferences of the labelers and the instructions given to the labelers based on which they were to rate the responses. Additionally, several other shortcomings exist such as prompt selection, assumption that human preferences are homogenous etc. Assembly AI's blog [2] highlights several other concerns that arise while using RLHF for tuning a large language model. Given its limitations, it is easy to conclude that outputs produced by LLMs trained by RLHF would have biases that are invisible to the prompter. It is this fine tuning that allows anti-natalism but does not allow suicide. This selective bias against certain controversial topics is ubiquitous and would be harmless unless such language models become the main source of our information. If such LLMs would be the sole source of information for a large group of people, it would naturally result in a very effective echo chamber.

Is there any indication of such LLMs becoming our source of information? Microsoft's search engine “Bing” being powered by GPT-4 is at least a first indication of where the industry is going towards. Other companies are following suit. Given that the industry is already moving in this direction, it is crucial to make sure that the reward models (or similar alignment mechanisms) that will be used by these LLMs will not be black boxes to the public. If unattended, a small subset of people will be able to tweak what kind of outputs these LLMs produce. Such power that these subsets of people have, historically seen, has been misused at every chance.

There is no reason to expect that such misuse will not happen in case of such LLMs. On the flip side, if this power is handed out to everyone, then we have a chatbot like Tay which is open to be corrupted but a subset of people. If inalienable human freedoms and rights are to survive, we have to make sure that information sources do not have inbuilt value filters. If such value filters are inevitable, then we need to make sure that these value filters are well known and agreed upon by independent stakeholders funded by the public. The lack of funding (and of unbiased funding) remains one of the main threats to the survival of the field of AI alignment. If AI alignment is not independently funded, researched, and enforced on LLMs, we are inviting invisible biases on LLM outputs.

© J C Varughese, April 2023

It is encouraging to see some conversation happening in the field of AI to operationalize a democratic approach towards value alignment in AI models. Although not without weaknesses, Deep Mind's paper [3] on value alignment considers how this might be best achieved. However, as start-ups and companies race to develop the latest technology, it is unclear how much effort will be actually invested into making sure that LLM are unbiased. Given the massive potential of LLMs to become the sole source of information for large groups of people, there exists a great risk of abuse. If this is not adequately taken care of by independent publicly funded parties, this will result in poisoning of our information sources.

References

1. https://arxiv.org/pdf/2203.02155.pdf?ref=assemblyai.com

2. https://www.assemblyai.com/blog/how-chatgpt-actually-works/

3. https://arxiv.org/abs/2001.09768

Written by Joshua Cherian Varughese (April 2023).

Joshua Cherian Varughese

Joshua is a roboticist who helps to build, test and commission robotic systems for industrial automation. He is currently working for a mobile robotics company in Linz, Austria. Previously, was a postdoctoral researcher at the Artificial Life Lab at the University of Graz (Austria). During his PhD at the Graz University of Technology (TU Graz), he developed a novel swarm paradigm to unify common swarm behaviours. His research interests lie in self-organization, swarm intelligence, swarm robotics and artificial intelligence among other things. He is also interested in understanding the social and ethical implications of AI and its applications.

© J C Varughese, April 2023