What you say is not what you mean: Large Language Models on Society and Environment

STS 4500 Prospectus

School of Engineering and Applied Science

Department of Computer Science

The University of Virginia, Charlottesville

Name: Neil Phan

Technical Advisor: N/A

STS Advisor: Alice Fox

Projected Graduation Date: 05/12/2024

Submission Date: 05/05/2023

What you say is not what you mean: Large Language Models on Society and Environment

Overview:

Throughout my time here at the University of Virginia, I have enjoyed taking classes that challenge me in learning the building blocks of computers through computer architecture, compilers, and other courses. One aspect that has been particularly interesting to discover is the connection of hardware accelerators to machine learning models, especially models such as ChatGPT from OpenAI and AlphaFold from DeepMind. For my STS project, we will have a comprehensive overview of the current state-of-the-art of large language models (LLMs), and the current implications it has on society and the environment. More specifically, the STS paper will cover the Actor Network Theory of OpenAI and its impact onto society and the environment. Readers will learn about the decisions that OpenAI has made into their creation of the GPT model series and ChatGPT, and what measures will be useful to ensure the safe creation of new LLMs for society's benefit.

Positionality:

There is lots of debate today on the direction of machine learning and artificial intelligence and where it will lead us into the future. With LLMs such as ChatGPT rising into the news and influencing our day-to-day lives, it comes to wonder what benefit can these models give with their cost? Researchers across academia and industry have come together to spend millions of dollars a singular model, consuming gigawatts of energy from grids across the world. The result are models that hopefully can help users answer common questions, create worldly stories, and much more, but is that cost necessarily worth the power? What can we do to reduce power consumption and train models effectively?

I hope to learn about what goes into the consumption of power and usability of machine learning models, particularly large language models, and debate whether or not models are being designed efficiently and effectively.

Projected Outcomes:

My goal is to explore the current societal effects of large language models in our day-to-day lives, observing how the public has changed in the past decade from AI technology. I will also explore the environmental impacts caused by machine learning model training given their large power usage to deploy large language models. By reducing power consumption, it'll also reduce negative impact on the environment that surrounds us.

Technical Project Description:

With my previous internship experience in data science and software engineering, along with a multitude of classes involving low level architecture and software design, this led me into my capstone project idea that I hope to work on throughout this and the next year: a compiler infrastructure for machine learning models that reliably and efficiently connects heterogeneous hardware to the high-level machine learning software libraries. There have been numerous research papers on the topic of hardware for machine learning, but a big bottleneck in adopting such hardware is creating a programming interface that current machine learning engineers can use to optimize their models. The goal of this project is to attempt to reduce the jargon in connecting the hardware to the software and lead to overall adoption that will accelerate machine learning training and research.

Preliminary Literature Review & Findings:

From my findings in the bibliography, most of the current research that is being done to contribute to LLMs is mainly focused on by either academic researchers or industrial researchers. For those who are directly working on the technical components of LLMs, there have been numerous resources that have described their cautious process in ensuring that explicit usage is minimized, such as OpenAI's alignment paper to reduce dangerous uses for LLMs when they were designing GPT-2 (Amodei et al., 2016). On the other hand, academic researchers without the economic motives have fought the idea that an intersection of social sciences with the development of artificial intelligence is necessary, gauging a need for diversity in developing these LLMs to maximize the safety of AI design (*A Guide to Solving Social Problems with Machine Learning*, n.d.; Hagendorff, 2020; Irving & Askell, 2019; Sloane & Moss, 2019). Algorithmic biases have also been a major concern with LLMs, with papers focused on the racial and gender biases that exist in mainstream LLMs (Abid et al., 2021; Bender et al., 2021; Bender & Friedman, 2018; Sun et al., 2019).

Environmentally, researchers have found that machine learning models are focused on optimizing speed which veers towards good signals (Huntingford et al., 2019; Strubell et al., 2020). However, energy-aware machine learning training has not been a big consideration that can open up more opportunity in energy-efficient machine learning training. This paper will also focus on the current

My work hopes to bridge the common understandings of researchers on developing safe, environmentally friendly large language models. It is imperative to develop an understanding of both sides that are supportive or against the current methodologies of designing LLMs.

STS Project Proposal:

STS is the research of how technology has formed and how it has impacted society. The discoveries in technology. Discoveries in machine learning have led to many uses in application, but the current overarching impacting of AI has been controversial. The given landscape of AI shows that it's a tool that has been created by society to help with many issues, but its implications on other prospects such as biases and the environment are concerning.

For my STS project, the main intent is to focus on the environmental and societal impacts of large language models on the landscape of today. To be more definitive, I hope to tackle the environmental impacts of training large ML models today and what effects they have had on the power grid and the environment itself. For the societal aspect, I hope to explore how large language models have infiltrated social media, ads, etc. to impact our political views, biases in race and gender, and how such data is being treated and used by both large companies and individuals for better or worse. Privacy is another big concern from recent news regarding scraping the entire internet to train the trillion of tokens the LLMs run on. Exploring the environmental concerns in LLMs connects back to my research on designing more efficient compilers for machine learning applications.

The main authors to follow are the industrial leaders that are leading the AI movement, the academic researchers that focus on AI safety and environmental machine learning research, and the public which uses AI products on a day-to-day basis. The work of the industry is imperative to follow given the recent effects it’s had on the public with releases of generative AI art, chat bots that can summarize information for you, and much more. The academic researchers provide an insightful perspective from many different fields that contribute to a complex and diverse overview of concerns with developing artificial intelligence.

To approach this piece, I will be using Actor Network Theory (ANT) to cover the impact of large language models. More importantly, I will consider the biggest actors that are contributing or are being affected by large language models, which would again be the industrial companies working on designing the LLMs, the academic researchers contributing towards the technical revolution of LLMs, other academic researchers focused on AI safety research and concerns in other fields, the public that ingests LLM products, and the environment that is training the machine learning models.

With ANT, it will help define a clearer connection between these major actors and other actors along with it. Through figuring out the positive and negative connections between these actors, it will help further the line on what we should be careful about when designing LLMs and what benefits we will be looking forward to in the future. This all connects back to STS paper on what are the current benefits and negatives from recent LLMs. To ensure that the ANT is in scope and can be analyzed in-depth, we will focus mainly on the LLMs designed by OpenAI such as the GPT series and ChatGPT.

I first plan to tackle the societal impact of large language models. To do so, I plan to read literature review of current AI safety topics that range from racial biases, gender biases, explicit content filtering, and more. There exist a lot of current research papers that tackle the different biases today, and is also tackled commonly in media from books, movies, TV shows, and many more. For the environmental aspect, this will be explicitly done through literature review and analysis and interviews. There are key individuals that work with hardware that have experience with power consumption with machine learning models that I hope I can interview with.

Barriers & Boons

From the curriculum at UVA, I have a lot of experience with the technical aspect of the research project, but I am on the lacking end when it comes to the thorough analysis to be done on the ethics of AI in society and the environment. To overcome this in this and next year, I plan to do more reading on what is being done for AI safety and AI’s impact on the environment and seeking advice from current UVA staff on what they have recommended to learn more about the topic. Another path I plan on taking is getting diverse opinions on the topic of AI and large language models itself by asking a diverse group of people of their current opinions on the state of technology and views.

References

A Guide to Solving Social Problems with Machine Learning. (n.d.). Retrieved March 13, 2023, from https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning

The authors include Jon Kleinberg, a computer science professor at Cornell University who specializes in large-scale networks and systems and their societal impacts; Jens Ludwig, a professor of law and public policy at the University of Chicago studying violence and urban poverty; and Sendhil Mullainathan, an economics professor at the University of Chicago who studies applying machine learning to study complex social problems. The article goes over how applying machine learning to social issues can help identify patterns that can deal with major problems of crime, domestic abuse, and many more issues. There are however issues that any scientist needs to be careful with when pursuing the field, such as data bias and ethical considerations. The authors recommend a framework to make ethical decisions by introducing a diverse range of stakeholders on the problem. The article is considerate in describing the overall idea and concerns but falls short in explicit cases. The source helps provide an academic perspective of positives of applying machine learning to societal problems.

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent Anti-Muslim Bias in Large Language Models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306. https://doi.org/10.1145/3461702.3462624

The authors are researchers at the Zou Group in Stanford University where Abubakar and Farooqi come from an Islamic background and are under the guidance of Zou and all of them come from a biomedical science and computer science background.
The paper examines the presence of anti-Muslim bias in large language models, such as GPT-3. The authors analyze the responses generated by these models to various prompts related to Islam and Muslims and find that the models exhibit a persistent bias against Muslims, perpetuating negative stereotypes and misinformation. The authors suggest that the bias is likely due to the lack of diverse representation in the training data and the need for better diversity and inclusion practices in the development of these models. They emphasize the importance of addressing bias and promoting diversity in the development and deployment of large language models to ensure that they are inclusive and equitable.

The paper does a good job of examining a specific bias in artificial intelligence and further shows the general idea that biases can come from the data. This is another unique perspective of the biases that can come from LLMs.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety* (arXiv:1606.06565). arXiv. http://arxiv.org/abs/1606.06565

The authors are computer science researchers at Stanford, UC Berkeley, Google Brain, and OpenAI who focus work on artificial intelligence research. The paper discusses several concrete problems related to ensuring the safety of artificial intelligence systems. They argue that as AI systems become more advanced and autonomous, they may pose significant risks to society if they are not designed with safety in mind. It highlights several specific challenges, such as the problem of avoiding negative side effects of AI systems, the problem of ensuring that AI systems do not behave in ways that are harmful to humans, and the problem of ensuring that AI systems are aligned with human values. The authors also discuss potential approaches to addressing these challenges, such as the use of value alignment techniques and the development of safe and transparent AI architectures. The paper emphasizes the need for continued research and development in AI safety to ensure that advanced AI systems are aligned with human values and do not pose a threat to society. This is relevant to the design of LLMs and involves researchers currently working on the same LLMs to ensure safety with releasing to the public.

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. https://doi.org/10.1162/tacl_a_00041

The authors include Emily Bender, a professor in linguistics at the University of Washington, and Batya Friedman, a professor in the Information School at the University of Washington. The paper presents a set of guidelines called Data Statements for Natural Language Processing that aim to promote transparency and reproducibility in the research and mitigate the risk of system bias. The guidelines encourage researchers to provide a detailed description of the data used to train and evaluate AI systems, including information on the demographics of the data and any potential sources of bias. The authors argue that by providing more information about the data used in their research, researchers can better understand the limitations and potential biases of their systems, and work to develop more inclusive and equitable NLP technologies. The paper also provides examples of how the Data Statements can be applied in practice and highlights the potential benefits of adopting these guidelines for the broader natural language processing research community. The paper connects to minimizing bias in natural language processing research and focusing on a diverse group of people when constructing systems such as LLMs.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? &#x1f99c; *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

The authors include Emily Bender from the previous paper; Timnit Gebru, a researcher from Stanford and Google Brain that advocates for diversity in tech; Angelina McMillan-Major, a graduate student researcher at the University of Washington focused on language technology; and Shmargaret Shmitchell, another advocate for diversity in tech that was part of Google’s Research team. The paper discusses the potential dangers of large language models such as GPT-3. The authors argue that LLMs have the ability to generate highly convincing fake text and have the potential to be used for malicious purposes, such as spreading misinformation and propaganda. It also highlights the ethical implications of the large energy consumption required to train and run LLMs, as well as the potential for reinforcing harmful biases in the data used to train these models. They suggest several recommendations to address these issues, including reducing the size of LLMs, promoting transparency and accountability in the development of these models, and promoting research on alternative approaches to language modeling that are more efficient and ethical. The paper covers the subject of misinformation which proves to be another important insight to creating safe LLMs. This directly connects to how LLMs need to be carefully designed for society.

García-Martín, E., Rodrigues, C. F., Riley, G., & Grahn, H. (2019). Estimation of energy consumption in machine learning. *Journal of Parallel and Distributed Computing*, *134*, 75–88. https://doi.org/10.1016/j.jpdc.2019.07.007

The authors come from both the Blekinge Institute of Technology in Sweden and the University of Manchester in the United Kingdom. The lead authors Eva García-Martín and Crefeda Faviola Rodrigues are both women researchers in computer science being supervised by professors Graham Riley and Håkan Grahn. The paper presents a methodology for estimating the energy consumption of machine learning (ML) algorithms running on different hardware platforms, such as CPUs, GPUs, and FPGAs. They propose a set of metrics to measure the energy consumption of different components of ML algorithms, including data transfer, computation, and memory access. It also provides a set of experimental results comparing the energy consumption of several popular ML algorithms running on different hardware platforms. The authors highlight the importance of estimating energy consumption in ML, as it can help optimize the design of ML algorithms and hardware systems to be more energy efficient. The paper provides a valuable contribution to the field of energy-aware machine learning, which aims to develop more sustainable and environmentally friendly ML systems. While a more technical approach to machine learning costs, the paper provides a closer guideline to measuring the environmental impacts of machine learning models that can contribute to the overall impacts of LLMs.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Thilo Hagendorff is a current AI researcher at the University of Stuttgart revolving around the ethics of AI, cognitive science, and human-animal studies. Hagendorff provides a comprehensive overview of the current state of AI ethics guidelines. They evaluate these guidelines based on their comprehensiveness, specificity, and practicality, and highlight the need for a more unified and coherent approach to AI ethics. They conclude that while the existing guidelines provide useful insights, they also suffer from various limitations, such as a lack of concrete recommendations and a failure to address power imbalances in the development and deployment of AI systems. The author suggests that future AI ethics guidelines should be based on a more holistic and interdisciplinary approach that considers the perspectives of diverse stakeholders and considers the broader societal implications of AI and balancing the technical components of artificial intelligence with the ethics of it. The author introduces interesting points of balance in research development of AI which connects close to treading forward with LLMs in environment and society.

Irving, G., & Askell, A. (2019). AI Safety Needs Social Scientists. *Distill*, *4*(2), e14. https://doi.org/10.23915/distill.00014

Amanda Askell is currently a philosophical researcher at the technical company Anthropic focusing on AI safety, while Geoffery Irving serves as a safety researcher for DeepMind. The paper argues that social scientists have a critical role to play in ensuring the safe and beneficial development of artificial intelligence. While technical AI safety research has made significant progress in recent years, they highlight the importance of understanding how AI systems will interact with society, as well as the ethical and political implications of their deployment. The authors suggest that social scientists can provide valuable insights into these issues and call for greater collaboration between AI researchers and social scientists to address the complex challenges posed by AI. Ultimately, they argue that a multidisciplinary approach to AI safety is necessary to ensure that AI is developed in a way that aligns with human values and preferences. The paper displays technical depth in alignment for AI safety and shows another benefit for understanding more how LLMs need the perspective of social scientists.

Sloane, M., & Moss, E. (2019). AI’s social sciences deficit. *Nature Machine Intelligence*, *1*(8), Article 8. https://doi.org/10.1038/s42256-019-0084-6

Mona Sloane is a sociology professor at New York University focused on AI design and policy, while Emanuel Moss is a PhD candidate at the CUNY Graduate Center focusing on the ethnography of data science. The paper examines the social sciences deficit in the field of artificial intelligence research and development. The author argues that the lack of input from social scientists in AI development has led to a narrow understanding of the social and cultural implications of AI systems. They discuss various examples of AI systems that have had negative social consequences, such as biased algorithms and surveillance technologies, and highlight the importance of social science expertise in addressing these issues. The author calls for greater collaboration between AI researchers and social scientists to ensure that AI systems are developed in a way that is socially responsible and ethical. To conclude, the paper highlights the need for a more interdisciplinary approach to AI research and development that incorporates social science perspectives, which provides another unique perspective on the intersection of social sciences to artificial intelligence. The demand for connecting social sciences to AI tools such as LLMs is important to cover.

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J. W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K., & Wang, J. (2019). *Release Strategies and the Social Impacts of Language Models* (arXiv:1908.09203). arXiv. https://doi.org/10.48550/arXiv.1908.09203

The authors contain a wide variety of researchers from OpenAI, Cornell University, Harvard University, CTEC, and Politwatch. The paper explores the social impacts of language models and the release strategies that can be used to mitigate potential harms. The authors show that language models have the potential to reinforce existing power structures and biases, perpetuate stereotypes, and enable harmful behaviors. They identify various release strategies, such as pre-release auditing, responsible disclosure, and community engagement, that can be used to identify and address potential harms before and after the release of language models. They also emphasize the importance of interdisciplinary collaboration and transparency to ensure that language models are developed and deployed in an ethical and socially responsible manner. Overall, the paper highlights the need for language model developers and researchers to consider the social impacts of their work and to adopt release strategies that prioritize ethical and responsible practices, which contributes closely to the need for researchers to take careful consideration in designing their LLMs for society.

Strubell, E., Ganesh, A., & McCallum, A. (2020). Energy and Policy Considerations for Modern Deep Learning Research. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(09), Article 09. https://doi.org/10.1609/aaai.v34i09.7123

The authors include Emma Strubell, a computer science professor at Carnegie Mellon University and part of Facebook Research; Anaya Ganesh, a master's student at the University of Massachusetts Amherst; and Andrew McCallum, a data mining professor at the University of Massachusetts Amherst. The paper examines the energy and environmental impacts of modern deep learning research and suggests ways to address these issues. The authors argue that the energy consumption of deep learning models has grown rapidly and has significant environmental implications, including carbon emissions. It discusses various strategies to reduce energy consumption in deep learning research, such as model compression, hardware optimization, and the use of renewable energy sources. They also suggest policy interventions that could incentivize researchers and developers to prioritize energy-efficient practices and technologies. The paper highlights the need for the deep learning community to consider the energy and environmental impacts of their work and to take action to reduce their carbon footprint, which connects closely with how LLMs need to also be designed with consideration for the environment.

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). *Mitigating Gender Bias in Natural Language Processing: Literature Review* (arXiv:1906.08976). arXiv. https://doi.org/10.48550/arXiv.1906.08976

The authors consist of two research groups from the University California Santa-Barbara and University of California Los Angeles. The paper provides a comprehensive survey of the literature on mitigating gender bias in natural language processing. The author examines different sources of gender bias, including biases in training data, word embeddings, and language models. The paper highlights various techniques for mitigating gender bias, such as debiasing algorithms, gender-specific word embeddings, and inclusive language models. The author also discusses the limitations of existing approaches and suggests directions for future research to further address gender bias in natural language processing. The paper provides a technical overview of how gender biases are perceived in natural language processing, showing ways that current LLMs and other AI models can discriminate by gender.