Undergraduate Thesis Prospectus

Entropi: An AI Sandbox for Product Development

(technical research project in Computer Science)

Bridging the Gap: The Initiative Towards Ethical AI Development

(sociotechnical research project)

by

Aneesh Vittal

October 27, 2023

technical project collaborators:

Manish Balamurugan

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

Aneesh Vittal

STS advisor: Peter Norton, Department of Engineering and Society

General research problem

How does the development of Artificial Intelligence (AI) affect people's sense of usefulness?

The concept of using emerging technology to increase efficiency and mitigate risks of manual work, despite consequences, has been realized across the globe. From mechanized looms that displaced textile workers, to assembly line robots that displaced factory workers, the social impact of technological progress is massive. A poll by the Pew Research Center revealed that 72% of those surveyed indicated concern over automation replacing their labor (Smith, 2017).

Work not only serves as a means of monetary pursuit but also provides a "vehicle for personal expression and offers us a means for personal definition" (Gini, 1998). Automation directly denies these opportunities for fulfillment to those it displaces. Recent developments in AI have redefined the scope of its utility: specifically, these developments blur the line between 'routine' tasks that are conducive to traditional automation and 'non-routine' tasks that do not adhere to a structured process (Susskind, 2023). Given the concern existing around automation and the ability of AI to cross the boundaries of tasks it can specialize in, further development has the potential to impact millions of people both economically and psychologically.

Entropi: an AI sandbox for product validation

How can LLMs and generative AI be used to simulate real-world conditions for the purpose of product development?

Entropi is an independent project I'm developing with Manish Balamurugan (mb2mcc@virginia.edu) in the Department of Computer Science. The technical advisor for this project is yet to be determined. The utility of generative AI has been explored in various contexts, from chatbots that aim to answer nearly any question to content generation tools that

produce results based on a textual description of what the user desires. The LLMs enabling such tools can take on specific roles as defined by contexts the user provides prior to any prompting. This role-playing ability powers applications such as AI therapists and customer service agents, but these often require manual prompting from the user.

In early 2023, researchers from Google DeepMind and Stanford University specifically explored fully autonomous interactions between role-playing generative AI agents. Agents representing various human profiles were introduced to a virtual environment and allowed to interact with one another through simulated conversations. More specifically, this research focused on "architectural and interaction patterns" that introduce human nuances, such as the ability to form opinions and memories over time, to the profiles "enabling believable simulations of human behavior" (Park et al., 2023).

Entropi applies these principles to host a simulated world of generative AI agents that discuss their opinions on products or ideas through autonomous interviews with an agent representing an interviewer working on behalf of the user. This would be especially beneficial to validate software products and services that tend to have large audiences spread across broad geographic regions. By automating interviews with a diverse set of agents representing the target audience for a product, users of Entropi can save countless hours spent manually sourcing and interviewing potential customers, along with the resources needed to conduct such product research. Entropi also enables underserved populations that may lack a network of individuals that can provide valuable feedback to drive a software startup.

An alternative to this approach uses Machine Learning (ML) models to classify blocks of text into classes such as 'negative', 'neutral', and 'positive' to indicate the author's sentiment. This method can be used to drive product development based on customer feedback such as

reviews, comments on social media, and testimonials. Software as a Service (SaaS) businesses such as Optimizely allow users to process and analyze data derived from product experiments. However, both tools require that users have pre-existing data and aren't suited to generate synthetic data that can serve as a basis for evaluation.

Entropi uses Python and Amazon Web Services (AWS) to host the backend of the system, along with a frontend built with ReactJS. The core conversational interactions between agents are powered by a simulation engine using fine-tuned GPT models through OpenAI's API and an initial set of profiles consisting of both original and AI-generated data. A demonstration of how Entropi works is available at https://entropi.app. The core limitation of this project lies in the ability to accurately simulate the very behaviors that determine one's sentiment towards products. Economic status, age, and lifestyle are only a handful of factors that influence people's purchasing decisions. Accurately conveying these factors, along with a myriad of others, through the simulated interviews is integral to producing a result that the user can rely on. Having developed a prototype, our focus is to test and develop the simulation engine towards producing hyper-realistic interactions that reveal actionable insights. This project has the potential to provide unfettered access to product experimentation and research to users of all backgrounds, technical or non-technical, supporting innovation across geographic boundaries.

Bridging the gap: the initiative towards ethical AI development

How do privacy advocacies, digital content creators, and media platforms organize to promote ethical development of AI?

Artificial Intelligence technology seems almost inescapable in 2023. While AI has largely been used on the enterprise level thus far, there has been a recent push towards consumer facing

AI products including chatbots and content generation tools that generate millions of dollars in revenue (Ludlow, 2023). Companies training these AI models often use data from external sources without following appropriate licensing procedures or gaining express consent (Blake, 2022). As these AI models are commercialized, digital content creators, privacy advocacies, and media platforms have increasingly voiced concerns over the use of their content to further AI companies' commercial goals, especially considering the lack of compensation for their contributions. Given the lack of current oversight on training AI models, how have groups organized to ensure interests of the owners of the data are protected?

With the rise of the internet since the early 2000s, stakeholders have tried to combat plagiarism across various industries. Napster, a popular online file sharing service, hosted troves of music, films, and other files that users would otherwise have to pay for. The popularization of Napster, along with torrent sites led to millions of dollars of lost revenue for the original creators of these files. Becker and Clement's research on the motivation driving users of these services found that users see themselves as members of a symbiotic community with a mutual goal. These users valued their freedom to access content that was normally locked behind pay-walls, or other authentication mechanisms, and were willing to face legal consequences for enabling such functionality. While this work answers the question of what factors drive the demand for illegally sourced content, it views this issue primarily through the perspective of the user and asserts that "the underlying motives of the users intensify" the development of mechanisms to subvert detection from copyright enforcement (Becker & Clement, 2006). Users may be one side of the equation: if there was no demand for platforms like Napster, resources would likely be dedicated to other projects. However, the responsibility of the platforms remains to be revealed.

Participants under this problem include three distinct categories with some variance in their interests: individual content creators, advocacy groups, and businesses. Paul Tremblay, an American author against copyright infringement in AI training has been vocal in challenging models like GPT and Meta's LLaMA that engage in "industrial-strength plagiari[sm]" (Saveri, 2023) of authors' work through legal means. Not only does the use of creators' work come without compensation, but it also damages future work opportunities. The Screen Actors Guild (SAG) has "prioritized the protection of [its] member performers against unauthorized use of their voice, likeness, and performances" as AI spreads into the film industry (SAG-AFTRA, 2023).

Similarly, organized advocacies like the Graphic Artists Guild promote a creator first approach and aim for graphic artists to have control over the use of their works pertaining to AI training (Blake, 2022). Their concerns lie with image generation models like DALL-E by OpenAI. These tools, having been trained on various artists works, have assumed the ability to express words and thoughts as images: a skill that defines artists creations and distinguishes one artist from another. This issue is not limited to the arts by any means. The Algorithmic Justice League (AJL) has been outspoken in developing bias conscious AI and encourages developers to be mindful of the societal harms they could cause by perpetuating discrimination in their work. The AJL places a large focus on the impact of facial recognition technology, in particular addressing the proven biases this technology has shown against marginalized groups (Algorithmic Justice League, n.d.).

Prominent companies, such as Reddit and X (Twitter), have also been outspoken on the use of data sourced from their platforms. Reddit's position on this topic is evident in their developer terms, stating that no "rights or licenses are granted" for the use of user content in

"training a machine learning model or artificial intelligence model" without express consent from the user (Reddit, 2023). Similarly, X prohibits any crawling or scraping of their services without expressed consent (Twitter, 2023). Reddit's role as an organized knowledgebase spanning various topics means they own large quantities of pre-labelled data that companies like OpenAI spend thousands on through their partner platforms like Remotasks. Labelled data is integral to producing viable natural language based AI products.

From legal action to educating the public, privacy advocacies, content creators, and businesses have been outspoken against the negative impacts further development of AI could have, especially if development is conducted without appropriate oversight. As AI becomes part of more facets of life, it becomes increasingly important to consider the impact on parties that would be adversely affected.

References

Algorithmic Justice League. (n.d.). *Spotlight - coded bias documentary*. Spotlight - Coded Bias Documentary. https://www.ajl.org/facial-recognition-technology

Becker, J. U., & Clement, M. (2006). Dynamics of illegal participation in peer-to-peer networks—why do people illegally share media files? *Journal of Media Economics*, *19*(1), 7–32. https://doi.org/10.1207/s15327736me1901_2

Blake, R. (2023, September 27). *Graphic artists guild issues statement of concern with AI image generators*. The Graphic Artist Guild. https://graphicartistsguild.org/graphic-artists-guild-issues-statement-of-concern-with-ai-image-generators/

Reddit. (2023, April 18). *Developer terms*. Reddit. https://www.redditinc.com/policies/developer-terms

Gini, A. (1998). Work, Identity and Self: How We Are Formed by the Work We Do. *Journal of Business Ethics*, *17*(7), 707–714.

Ludlow, E. (2023, August 30). *OpenAI nears $1 billion of annual sales as CHATGPT takes off*. Bloomberg.com. https://www.bloomberg.com/news/articles/2023-08-30/openai-nears-1-billion-of-annual-sales-as-chatgpt-takes-off

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. https://doi.org/10.1145/3586183.3606763

SAG-AFTRA. (2023, March 17). *AFTRA statement on the use of artificial intelligence and digital doubles in media and entertainment*. SAG. https://www.sagaftra.org/sag-aftra-statement-use-artificial-intelligence-and-digital-doubles-media-and-entertainment#:~:text=We%20will%20continue%20to%20negotiate,to%20aiquestions%40sagaftra.org.

Saveri, J. (2023, June 28). *LLM litigation*. LLM litigation · Joseph Saveri Law Firm & Matthew Butterick. https://llmlitigation.com/

Smith, A. (2017, October 4). *2. Americans' attitudes toward a future in which robots and computers can do many human jobs*. Pew Research Center: Internet, Science & Tech. https://www.pewresearch.org/internet/2017/10/04/americans-attitudes-toward-a-future-in-which-robots-and-computers-can-do-many-human-jobs/

Susskind, D. (2023, January). *Work and meaning in the age of ai - brookings*. Brookings. https://www.brookings.edu/wp-content/uploads/2023/01/Work-and-meaning-in-the-age-of-AI_Final.pdf

Twitter. (2023, September 27). *X terms of service*. Twitter. https://twitter.com/en/tos