Bridging the Gap: The Initiative Towards Ethical AI Development

An STS Research Paper
presented to the faculty of the
School of Engineering and Applied Science
University of Virginia

by

Aneesh Vittal

March 14, 2024

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.

Aneesh Vittal

STS Advisor: Peter Norton

Bridging the Gap: The Initiative Towards Ethical AI Development

Though enormously useful, AI threatens employment in many sectors, and compromises the integrity of intellectual property. Privacy advocates, content creators, and media platforms have objected to AI companies' use of their intellectual property without compensation. While most experts agree that AI requires regulation, AI optimists argue that strict regulation threatens to deprive society of some of the substantial benefits that AI has to offer. Proponents of ethical AI, however, argue that strict regulation of AI is necessary to protect human rights, social equity, and intellectual property.

Review of Research

From 1999 to 2001 Napster, an online filesharing service, hosted troves of music and films. The free files cost rightsholders millions of dollars in lost revenue. Becker and Clement (2006) found that users of such services see themselves as members of a community with a mutual goal. Users valued their freedom to access paywalled content and were willing to face the legal risks. While this work highlights factors driving the demand for illegally sourced content, it views this issue primarily through the user's perspective and asserts that "the underlying motives of the users intensify" the development of mechanisms to subvert 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.

The sustainability movement requires a similar consideration from automakers. For example, Aladashvili (2024) finds that producers of Battery Powered Electric Vehicles (BEVs) remain responsible for ensuring materials are sourced ethically. Consumer demand alone does not justify disregard for the common good. Similarly, AI companies have a duty to evaluate their practices to ensure positive outcomes for society. Standing at a pivotal moment in AI development, we must make meaningful design decisions. While AI has great utility as a productivity aid, stakeholders in AI companies and governments must consider the risk their policies pose to people, their livelihoods, and their rights.

A Bleak Future

The term Artificial Intelligence connotes a robotic know-it-all with self sufficient knowledge, but that's far from the truth. Humans are often the source of the data powering a model's intelligence. In fact, companies like Open AI and Scale AI rely on outsourcing the labor required to label their training data to Remotasks, or similar gig-work style platforms. Remotasks allows users to complete tasks like correcting lines of code or ranking text based responses generated by AI. Users of Remotasks are diverse, with many concentrated in countries like Brazil and the Philippines that lack career opportunities. Even then, the reality of working for Remotasks is dark. "Extremely low rates, routinely delayed or withheld payments", and poor labor standards were all common complaints from those interviewed among the 10,000 or so Filipinos working for the platform (Cabato & Tan, 2023). With AI threatening alternative careers and driving the low pay environment they've become accustomed to, workers may have nowhere else to turn.

Around a third of white collar workers fear being replaced by AI according to Asana's State of AI Work (2023) report. This concern is not baseless. In fact, workers are already facing consequences of AI adoption. Olivia Lipkin, a copywriter from San Francisco testifies to feeling "insecure and anxious that [ChatGPT] would replace" (Verma & De Vynck, 2023) her. Months later, Lipkin's role was replaced by ChatGPT. Forcing humans to compete with AI in the labor market is a burden that we are yet to see the effects of. While business efficiency drives companies to consider AI's cost over employing human capital for the same tasks, the opportunity to preserve labor markets lies in training employees to work alongside AI, without the fear of being replaced. To this end, The White House's 2022 report advocates for the "need to redesign" job functions across roles that could be impacted by AI. This report also references a study by Bessen et. al. (2019) that underscores the financial impact that AI automation has on workers' income, with sparing reassurances from safety nets like unemployment benefits.

These findings provide a window into a future of reckless automation with people facing adverse effects in the name of optimizing efficiency above all else. While humanity has faced similar workforce disruption in the past, periods like the Industrial Revolution presented a diverse set of career opportunities for workers to apply their existing skillset. However, AI's ability to automate non routine tasks means it not only threatens career opportunities, but also the scope of the job market. As workers upskill to pivot to careers spared by AI, many will be left in the dark with financial and physical barriers to obtaining required skills. With AI development rapidly advancing, it is imperative to consider the burden it places on workers and availability of a fair, competitive, and vast labor market.

Concerns With AI Training Procedures

Proponents of ethical AI condemn AI companies for pursuing "first-to-market" advantages at any cost. The race to deploy high performing AI models has driven companies like OpenAI to resort to rapidly sourcing data, often obtaining content from businesses and individual creators without license or attribution. 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.

Plagiarism

In December 2023, The New York Times voiced similar concerns in their lawsuit against Open AI. The Times challenged that AI industry giants like Open AI and Meta’s use of their authors’ content has resulted in competing sources of news (Grynbaum & Mac, 2023). In response, a spokesperson from Open AI encouraged the Times to adapt their business to “new revenue models”, effectively imposing adoption of AI on the news industry. Perplexity AI, a startup out of San Francisco, can provide concise news summaries powered by data scraped directly from news organizations’ websites. If newsreaders can obtain the information through more accessible methods like Perplexity, media organizations may have to become less reliant on subscriptions or advertising on their platforms and instead be forced to form licensing agreements with AI companies.

This concern isn’t limited to news media by any means. The Screen Actors Guild (SAG), representing actors globally, 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) via collective bargaining. 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 and generation models like DALL-E and Sora 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 distinguishes artists' creations from others'.

Ultimately, these groups stand united in ensuring the security of intellectual property owned by those they represent and taking action to protect themselves. While contracts and licensing agreements help creators adapt to a future with AI, limiting monetization of creative outlets like art and writing make them less lucrative. The legal landscape surrounding AI-generated content is still evolving, and current intellectual property laws may not adequately address the unique challenges posed by AI. As Samuelson (2023) notes, "AI systems can generate outputs that are similar to, but not identical to, the works on which they were trained, raising questions about the extent to which such outputs infringe copyrights" (p. 15). This legal ambiguity creates uncertainty for both AI companies and content creators, highlighting the need for updated legislation and clear guidelines to protect intellectual property rights in the age of AI. Regulation and enforcement of intellectual property laws will require further scrutiny to reflect the risks this technology poses.

Privacy

The rapid advancement of AI has brought forth critical concerns about privacy and data practices. AI systems powering social media algorithms rely heavily on vast amounts of personal data to train their models and make content recommendations. Companies track user behavior in order to target users with content and advertisements. However, this data is often collected and used without clear consent from individuals, raising serious privacy issues. Companies have been accused of exploiting user data to further their commercial goals, often without providing fair compensation or transparency about their practices.

Regulators have scrutinized changes companies made to their privacy policies, allowing them to employ user data to train AI models. The Federal Trade Commission (FTC) showed its commitment to rooting out companies that quietly modify terms of service in a policy recommendation by Fondrie-Teitler (2024). The report highlights the impact of unethical data collection with companies like 1Health that handle our most sensitive data, like DNA samples, being party to such practices. Pearce (2021) yields further perspective on privacy concerns from the perspective of ISACA, an IT governance organization. AI presents further privacy risks than just the initial exposure of data: personal data can persist and even outlive their sources post model training (Pearce, 2021). While users are often presented with options to opt out of data collection, once a model has been trained, it becomes difficult to remove just a sliver of the data, rendering the opt out procedure obsolete.

Despite the privacy risks, AI also has the potential to enhance privacy protections when developed and deployed responsibly. Privacy-by-design principles can embed privacy measures into AI systems from the ground up, such as data minimization and anonymization techniques. To this tune, Kerry (2020), a researcher from the Brookings Institution, advocates for the United States to adopt data privacy regulation that parallels the European Union's (EU) GDPR. With controls in place to limit intrusive data collection, putting users' interests first leads to a more desirable outcome for all parties.

In addition to concerns about data collection and use, AI also raises the specter of increased surveillance and monitoring. As AI systems become more sophisticated, they can be used to analyze vast amounts of data from various sources, including social media, facial recognition systems, and IoT devices, to track individuals' behaviors and preferences (Zuboff, 2019). This pervasive surveillance can have a chilling effect on free speech and individual autonomy. To mitigate these risks, robust data governance frameworks and privacy-preserving technologies, such as differential privacy and federated learning, should be implemented to ensure that AI systems are developed and deployed in a manner that respects individual privacy rights (Floridi, 2019).

Equity

Ensuring equity and fairness in AI training and development is crucial to prevent the perpetuation of societal biases and discrimination. AI models learn from the data they are trained on, so it is essential that this training data is diverse, representative, and free from biases. However, achieving this can be challenging as historical data often reflects existing inequities and prejudices. Amazon's AI hiring project in 2014 showed how these inequities are ingrained in biased datasets. The research team trained their model using resumes aggregated over 10 years. The model selected male applicants as the most ideal candidates, reflecting the patterns of the training data (Iriondo, 2018). If the data being supplied to models is biased or flawed, the outputs will reflect the same.

To mitigate these issues, AI developers must take proactive steps to identify and remove biases in training datasets. This involves carefully auditing data using techniques like adversarial debiasing to reduce discrimination, and continuously monitoring AI models for fairness. Companies like Optum are already using these strategies and report promising results. Optum's services aimed to allow healthcare providers to evaluate patients' long term needs. Their data and training practices, however, resulted in reinforced racial and financial biases in their model's outputs. Researchers re-evaluated the source of the data they used, taking care to represent data proportionally, and found an 84% reduction in the biases it produced (Paul, 2019).

Further, promoting diversity within AI development teams can bring varied perspectives to identifying and addressing bias. Diverse perspectives are crucial to securing ‘emotional intelligence’ as a core tenet of companies’ design process. Intentional design will lead to better accessibility and less bias as prioritizing the human impact of their work will allow teams to identify and address biases and inequity in all facets of their work (Shastri, 2020). Explainable AI (XAI) is another important tool for promoting equity and fairness in AI systems. XAI techniques aim to make AI models more transparent and interpretable, allowing developers and users to understand how the models arrive at their decisions (Adadi & Berrada, 2018). By providing insights into the factors that influence AI outputs, XAI can help identify and mitigate biases, as well as facilitate accountability and trust in AI systems. Incorporating XAI principles into the AI development process can contribute to more equitable outcomes and ensure that AI systems are used in a responsible and ethical manner. Transparency around AI training processes and datasets via public audits can also help build trust in the equity of these systems.

Regulation and Impacts

Beyond its impact on jobs and privacy, AI has far-reaching societal implications that demand careful consideration. One major concern is the potential for AI to exacerbate existing inequalities. Some argue that benefits of AI will accrue disproportionately to large tech companies and highly skilled workers, leaving others behind, while others argue for optimism in new opportunities arising from the AI boom. Policymakers must strike a delicate balance between enabling innovation and protecting societal interests. Overly restrictive regulations could stifle AI progress and hinder competitiveness, while under-regulation leaves the door open for misuse.

The United States has already begun the roll out of new legislation and regulation in this space. President Biden’s 2023 executive order focused on establishing “new standards for AI safety and security” that promote transparency in the development process, safety standards, and mitigate uses of AI for fraud (The White House, 2023). While the effects of this order are yet to be seen, it shows the US’ commitment to acting quickly to prevent misuse. In his Senate testimony, OpenAI CEO Sam Altman called for the establishment of a new licensing agency to oversee compliance and safety standards for AI models above a certain capability threshold. Altman proposed that this agency would create a set list of safety standards and require companies to engage in independent audits (Allyn, 2023). While his engagement with policymakers earned some praise (Feiner, 2023), critics have expressed skepticism about Altman's intentions. The proposed licensing process would heavily benefit OpenAI and entrench its position as a leading AI company (Allyn, 2023). When pressed on specifics, Altman declined to endorse suggestions such as transparency requirements for AI training data or commitments not to train models on copyrighted material.

If key stakeholders in AI cannot be trusted to help develop policy, governments should look to peer nations and advocacy groups for guidance. In late 2023, The EU came to an agreement on the AI Act that imposes sweeping regulations. The legislation includes “transparency requirements [for] chatbots and software that create manipulated images” and restricts governments’ use of AI for facial recognition and law enforcement. Violations of this act would result in “fines of up to 7 percent of global sales”, further underscoring the EU’s priority to ensure responsible adoption of this technology (Satariano, 2023). Enforcement will be crucial to realizing the positive impact of these legislations, holding companies accountable throughout the process. Effective regulation of AI will require close collaboration between governments and industry. Public-private partnerships can play a key role in promoting responsible AI development. For example, the Partnership on AI is a multi-stakeholder organization that brings together leading technology companies, civil society groups, and academic institutions to develop best practices and tools for responsible AI (Partnership on AI, n.d.). Similarly, the U.S. National Institute of Standards and Technology (NIST) has launched a collaborative effort with industry to develop a voluntary AI risk management framework. "The framework aims to help organizations identify and manage risks related to AI systems, including issues of bias, transparency, and accountability" (NIST, 2021, para. 2). By working together, governments and industry can ensure that AI is developed in a way that benefits society while mitigating potential harms.

To ensure the ethical development and use of AI, governments should first establish risk-based regulation that defines the highest risk uses of AI and focuses oversight on specific contexts rather than the technology itself (Feiner, 2023). Second, implementing transparency and disclosure requirements, ensuring that consumers know when they are interacting with an AI system and that companies disclose training data and model performance (Allyn, 2023). Third, protecting intellectual property by clarifying that using copyrighted content to train AI models is not "fair use" and exploring licensing regimes similar to music rights clearinghouses (Tracy, 2023). Finally, requiring independent audits and monitoring of AI models, especially those for high-risk applications, with results made publicly available (Feiner, 2023). Those impacted by AI adoption attest to these policies and hope governments will take action.

Conclusion

Ethical AI development is integral to protecting the future of opportunity across the world. People are already facing the consequences of companies chasing innovation at all costs and are advocating for themselves. Society’s concerns are not unfounded considering the bleak consequences of mass unemployment and an idle society with potential for lasting impacts to generations to come. These effects will be imperative to consider as adoption of AI scales into facets of daily life with those impacted leading the call for a mindful and ethical future. This means prioritizing transparency, accountability, fairness, and respect for privacy in the design and use of AI systems, and overall more intentional design. It requires active collaboration among all stakeholders—policymakers, industry leaders, researchers, civil society groups, and the public—to shape the trajectory of AI in service of the common good. Time will tell how effective existing regulation will be, but governments and stakeholders must get used to the idea of revisiting, re-evaluating, and updating legislation as circumstances change. Listening to constituents and the concerns expressed by advocacies allows for politicians to create legislation that captures their best interests. This is the only way to ensure effective enforcement and policy guidance that evolve alongside technology.

References

Adadi, A., & Berrada, M. (2018). *Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)*. IEEE Access, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Aladashvili, G.A. (2024). *EXTENDED PRODUCER OBLIGATIONS – AS A TOOL FOR REDUCING NEGATIVE ENVIRONMENTAL IMPACT.* The New Economist.

Allyn, B. (2023, May 16). *OpenAI CEO Sam Altman testifies before the Senate Judiciary Committee [Radio broadcast]*. NPR. https://www.npr.org/2023/05/16/1177780982/openai-ceo-sam-altman-testifies-before-the-senate-judiciary-committee

Asana. (2023). *State of AI at Work*. https://asana.com/work-innovation-lab/wp-content/uploads/2023/08/The-State-of-AI-at-Work.pdf

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

Bessen, J. E., Goos, M., Salomons, A., & Van den Berge, W. (2019). *Automatic reaction - what happens to workers at firms that automate?* SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3328877

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/

Cabato, R. & Tan, R. (2023, August 28). *Behind the AI boom, an army of overseas workers in 'digital sweatshops'.* The Washington Post. https://www.washingtonpost.com/world/2023/08/28/scale-ai-remotasks-philippines-artificial-intelligence/

Feiner, L. (2023, May 16). *OpenAI CEO Sam Altman urges U.S. to regulate A.I. in Senate testimony*. CNBC. https://www.cnbc.com/2023/05/16/openai-ceo-sam-altman-urges-us-to-regulate-ai-in-senate-testimony.html

Floridi, L. (2019). *Establishing the rules for building trustworthy AI.* Nature Machine Intelligence, 1(6), 261-262. https://doi.org/10.1038/s42256-019-0055-y

Fondrie-Teitler, S. (2024, February 13). *Ai (and other) companies: Quietly changing your terms of service could be unfair or deceptive*. Federal Trade Commission. https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/02/ai-other-companies-quietly-changing-your-terms-service-could-be-unfair-or-deceptive

Grynbaum, M. & Mac, R. (2023, December 27). *The Times sues OpenAI and Microsoft over A.I. use of copyrighted work*. The New York Times. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

Iriondo, R. (2018). *Amazon scraps secret AI recruiting engine that showed biases against women - machine learning - CMU - Carnegie Mellon University*. Machine Learning | Carnegie Mellon University. https://www.ml.cmu.edu/news/news-archive/2016-2020/2018/october/amazon-scraps-secret-artificial-intelligence-recruiting-engine-that-showed-biases-against-women.html

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

National Institute of Standards and Technology. (2021, July 29). *NIST launches artificial intelligence risk management framework.* https://www.nist.gov/news-events/news/2021/07/nist-launches-artificial-intelligence-risk-management-framework

Partnership on AI. (n.d.). *About us*. https://partnershiponai.org/about/

Paul, K. (2019, October 25). *Healthcare algorithm used across America has dramatic racial biases*. The Guardian. https://www.theguardian.com/society/2019/oct/25/healthcare-algorithm-racial-biases-optum

Pearce, G. (2021, May 28). *Beware the privacy violations in Artificial Intelligence Applications.* ISACA. https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2021/beware-the-privacy-violations-in-artificial-intelligence-applications

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%40 sagaftra.org.

Samuelson, P. (2023). *Copyright and AI-generated works*. Communications of the ACM, 66(4), 14-16. https://doi.org/10.1145/3579123

Satariano, A. (2023, December 8). *E.U. agrees on landmark artificial intelligence rules.* The New York Times. https://www.nytimes.com/2023/12/08/technology/eu-ai-act-regulation.html

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

Shastri, A. (2020, July 1). *Diverse teams build better AI. here's why*. Forbes. https://www.forbes.com/sites/arunshastri/2020/07/01/diverse-teams-build-better-ai-heres-why/?sh=2d9c9b4477b3

The White House. (2022, December). *The impact of artificial intelligence on the future ...* https://www.whitehouse.gov/wp-content/uploads/2022/12/TTC-EC-CEA-AI-Report-12052022-1.pdf

The White House. (2023, October 30). *Fact sheet: President Biden issues executive order on safe, secure, and trustworthy artificial intelligence.* The White House. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

Tracy, R. (2023, June 13). *News Industry Seeks Tougher Rules for AI Use of Its Content*. The Wall Street Journal. https://www.wsj.com/articles/news-industry-seeks-tougher-rules-for-ai-use-of-its-content-e8d1f9e4

Verma, P. & De Vynck, G. (2023, June 02). *CHATGPT took their jobs. now they're dog walkers and HVAC techs*. The Washington Post. https://www.washingtonpost.com/technology/2023/06/02/ai-taking-jobs/

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.