GENRATIVE ADVERSARIAL NETWORKS AND DIFFUSION IN SYNTHETIC IMAGES

ADDRESSING THE ETHICS OF AI INTERACTIONS

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Luke Benham

December 2, 2022

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.

ADVISORS

Catherine Baritaud, Department of Engineering and Society

Briana Morrison, Department of Computer Science

The development of Artificial Intelligence will likely be the most important technology of the coming decades. The recent advances in image generation models such as DALL-E 2 and Stability Diffusion have demonstrated AI programs are capable of creative work previously thought to be solely in the domain of human artists. This has accelerated the belief that all human abilities will eventually be outperformed by AI leading to Artificial General Intelligence that exceeds human capabilities across all domains. The State of the Art Technical Report will focus on image generation models and their use of Generative Adversarial Networks and diffusion. These techniques have led to the most advanced synthetic images and diffusion is promising for the next iteration of image generators, as well extending into the new creations in 3D object and video generation.

The STS Research Project is loosely coupled with the technical report as it explores the ethics of AI including but not limited to image generation. The development of AI will affect society both in its new capabilities and also in the humans that it will replace. As the development of AI is becoming so rapid and difficult to predict, the ethical principles for integrating AI into society should be established early. Actor Network Theory will be used to evaluate the interactions between the relevant entities. Image generation models are a useful case study as they are currently disrupting art and creative industries and the ethics of their development remain unclear.

The creation of this STS Research Paper will be advised by Professor Catherine Baritaud and Professor Briana Morrison will be the advisor for the Technical Report. The timeline of the Technical Report and STS Research Paper will run in parallel with three major phases each lasting approximately a month. The first phase will be dedicated to research, development, and outlining both papers. The second part will contain the majority of the drafting process until a rough draft is created. The final month will be spent revising and connecting the entire thesis.

GENRATIVE ADVERSARIAL NETWORKS AND DIFFUSION IN SYNTHETIC IMAGES

The trend in the development of Artificial Intelligence is towards machine learning involving deep neural networks that are increasingly opaque to human understanding. Generative Adversarial Networks exemplify this issue as they use unsupervised training and the discriminator model iteratively creates the utility function for the generation model. Diffusion extends this issue as the training involves reversing random Gaussian noise which leads to a level of randomness that is near impossible to decode once the model is trained. The lack of transparency leads to issues detecting bias, bugs, or unintentional outputs from the models. Despite these potential downsides, image generators have demonstrated incredible abilities that will transform the online landscape.

The state of the art technical report aims to collect and project the recent progress in image generation from GANs and diffusion models using resources from both academic and applied settings. The field is developing so rapidly that academic papers often focus on the mathematical basis for machine learning while the documentation from the organizations creating the most powerful models can contain practical insight into the effects of these systems. The public research and algorithms often provide the foundation for new models, sometimes intentionally by the authors. The performance of image generators in common topics such as colorization, inpaining, uncropping, and JPEG restoration can be measured against the benchmarks set by academics (Saharia et al., 2022, p. 4). Models created with GANs are highly capable in digesting complex data with high dimensionality (Hong et al., 2020, p. 1). They are able to learn from this data due to their unique structure shown in Figure 1 (Hong et al., 2020, p. 4). The benefits of GAN models are not only in their outputs themselves, but also their ability to create high quality data for other important applications including biomedical images (Canas et al., 2018, p. 2).

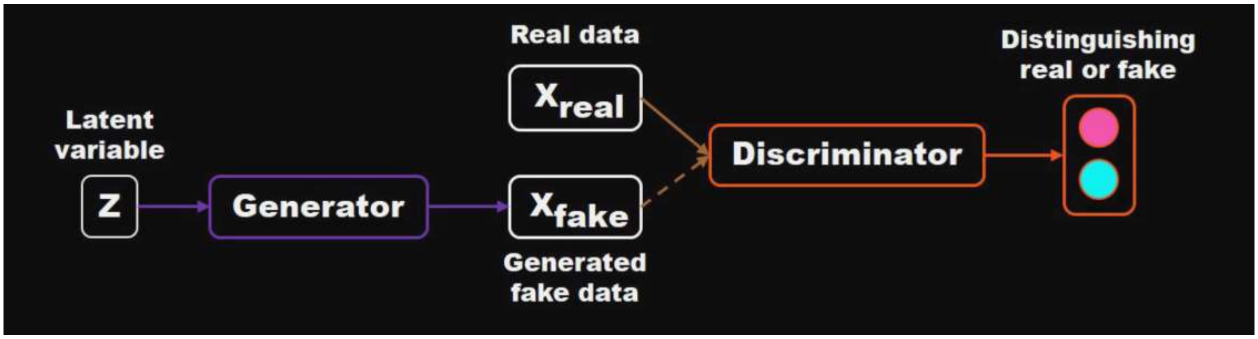

Figure 1: Diagram of the structure of a Generative Adversarial Network model. This figure demonstrates the pipeline through which the GAN uses data for self-improvement. (Hong et al., 2020, p. 4)

The complexity of the cutting edge models will continue to rise as a function of both the more complicated algorithms made possible by increased computing power and the inherent complexity of new forms of data. For example, the use of GANs for video generation may use 3D convolutional neural networks instead of the 2D networks seen in image GANs to account for the time dimension (Aldausari et al,, 2022, p. 8). Similarly diffusion models have high possible degrees of complexity across domains including some relatively simpler domains like denoising creating non-local filters that increase efficacy across textured or complex images (Qiao et al., 2017, p.7). The ultimate goal of AI to produce superhuman level abilities is not only being approached through complexity but also by fusing multimodal data together to process combined image and text data (Luo et al, 2021, p. 1). By navigating the intricacies of the current image generation research landscape, the state of the art technical report hopes to map the frontiers of the best AI models. The scholarly article will both evaluate the techniques currently used and estimate the possible future domains and abilities in which image models will make transformative achievements.

ADDRESSING THE ETHICS OF AI INTERACTIONS

Artificial Intelligence creates the urgency to formalize and codify human morality, a goal that has evaded philosophers for centuries. Ethics must be established in advance as "it will take time for formal law and regulations to catch up with the technological developments." (Kuleshov et al., 2020, p. 2). The need to have clear and logically consistent principles is becoming more of a necessity as AI exerts greater influence over society. There are basic components of an ethical system that are generally agreed upon but have not been turned into regulation. If ethical design is not implemented with coming AI systems, the issues of non-compliance and bias may be continued or exacerbated by unsupervised learning algorithms. This could happen either through biased data that fails to reflect the proper human diversity or by locking in wrong ethical norms. It is currently debated whether this value lock in will result from conscious machines or if less sophisticated AI are still capable of outperforming humanity without our defining consciousness (O'Lemmon, 2020, p. 1). While this has serious moral implications on the value of a given AI system it still seems likely that the majority of future utility lies in the degree to which humans and their values are protected.

The previous work on AI ethics has generally either remained too technical in the attempt to solve specific problems with individual AI systems or has taken a sociological approach that focuses on the human effects without properly accounting for the agency of advanced programs. Progress could be made by using a practical and utilitarian foundation to evaluate AI through applied ethics as demonstrated in Figure 2 on page 5.

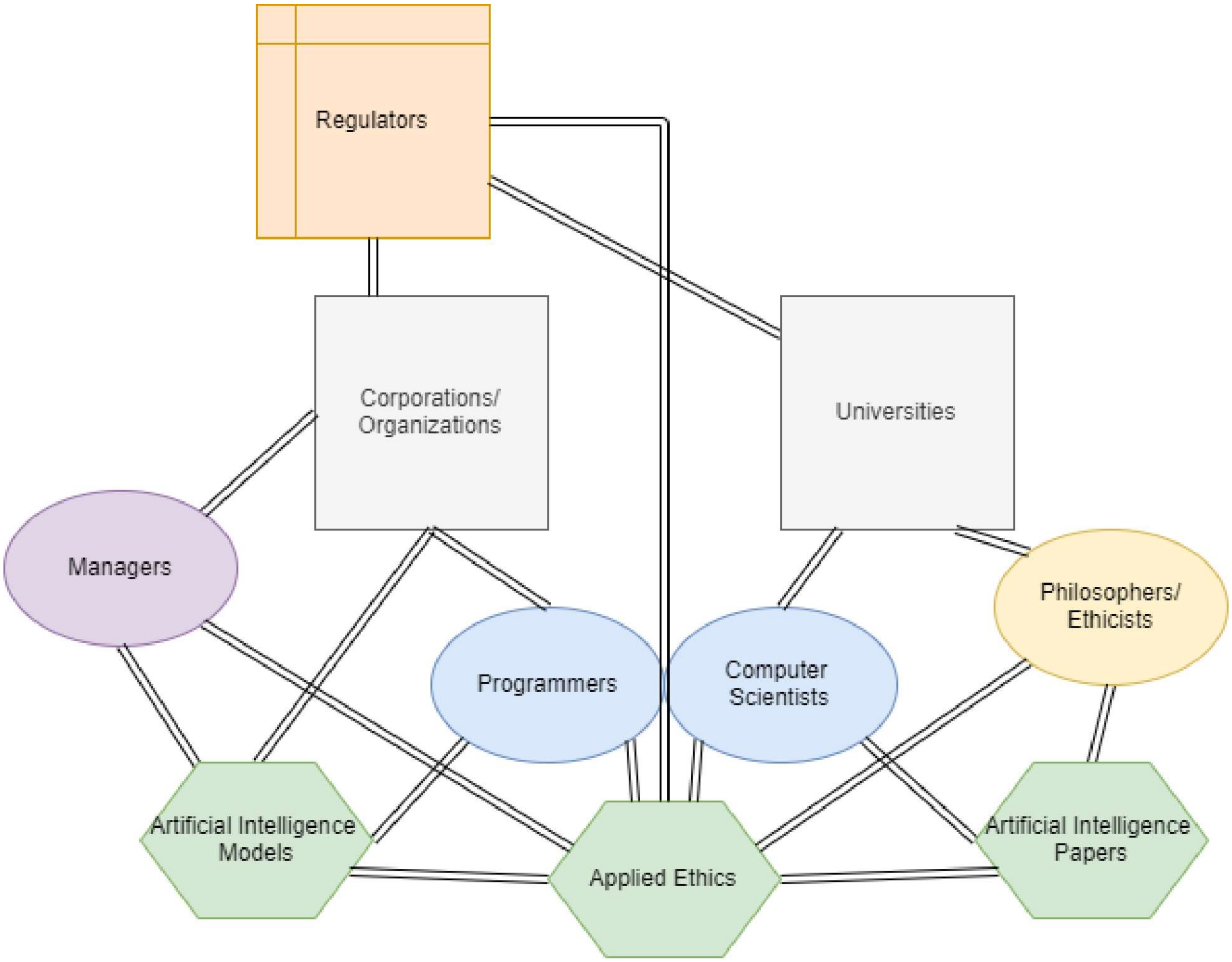

Figure 2: A diagram of the groups that contribute to the Applied Ethics of AI development. It illuminates the numerous inputs that lead to the ethical frameworks for the field. (Benham, 2022)

Actor Network Theory (ANT) can also be used to evaluate the interactions of AI with the sociotechnical system (Kaartemo & Hekkua, 2018, p. 4). By evaluating AI through the lens of ANT, the value co-creation of AI in concert with human actors can be dissected into four themes. Co-creation occurs across field advancement, supporting service providers, enabling resource integration, and supporting beneficiary well-being (Kaartemo & Hekkua, 2018, p. 5). These themes touch on the agency of the AI in a network. The ANT tenet that "humans and technologies cannot be fully separated" should be emphasized as AI systems move from being tools towards being partners in human endeavors (Bengtsson, 2018, p. 7). The building of a model using ANT will also require the ideas of both social constructivism and technological determinism on how society is shaped by technology (Matthews, 2020, p. 1). Figure 3 shows how technological determinism would blur into social constructivism as AI is increasingly considered a social actor. Using this framework is important to understanding the tech-user relationship, especially in sensitive fields such as the intelligence community (Vogel, 2021, p. 1). The STS research paper will therefore address the ethical risks that AI will pose to humans through ANT by considering their interactions with humans and other advanced technology.

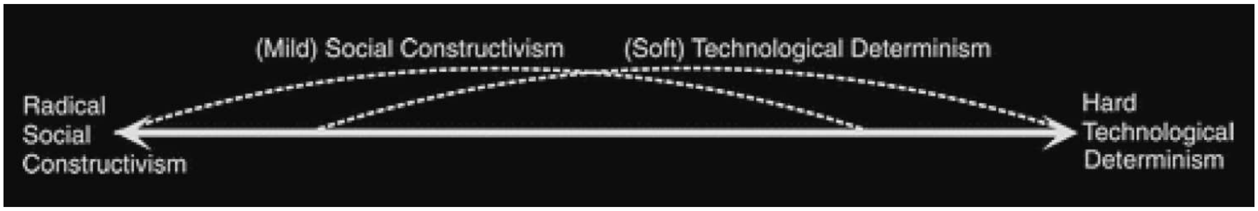
(Mild) Social Constructivism
(Soft) Technological Determinism
Radical Social Constructivism
Hard Technological Determinism

Figure 3: Spectrum of beliefs on the relative importance of technology and society on progress. It demonstrates the overlap of the given viewpoints. (Dafoe, 2015, p. 4)

EXPONENTIAL CHANGES

Human cognition is notoriously bad at properly estimating exponential trends. As the influence of AI grows exponentially, the regulation and planning surrounding its development must change to become more proactive. This structure is necessary to deal with the ethical concerns of AI because the feedback loop that normally drives socio-technical change will become too quick for society to meaningfully adapt to the consequences of unaligned AI. The creation of an Artificial General Intelligence exceeding all human abilities is drawing increasingly near in expert and crowdsourced predictions. According to Metaculus, a site that hosts prediction markets where real money is bet on future events, AGI is expected in 2040 as shown by Figure 4 on page 7.

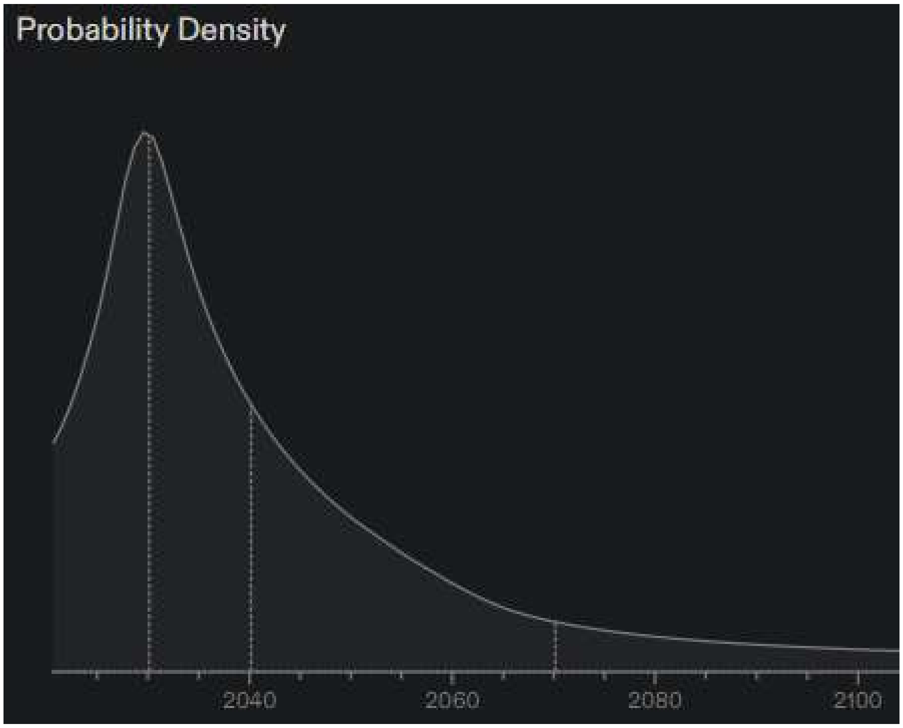

Figure 4: Graph showing the probability density of the date that Artificial General Intelligence is announced. The graph has the average prediction at 2040 with a long right tail. (Barnett, 2020)

The rapid progress of technology will be difficult for humanity to control. The stabilization of a technological artifact may break down as the acceleration of scientific progress exceeds the corrective steering of societal norms. A sufficiently powerful AI system may not be able to be aligned after it is created, so humanity must be prepared before that day arrives.

REFERENCES

Aldausari, N., Sowmya, A., Marcus, N., & Mohammadi, G. (2022). Video generative adversarial networks: a review. *ACM Computing Surveys, 55*(2)*,* 1–25. https://doi.org/10.1145/3487891

Barnett, M. (2020, August 23). *When will the first general AI system be devised, tested, and publicly announced?* Date of Artificial General Intelligence. Retrieved November 1, 2022, from https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/

Bengtsson, S. (2018). Ethics Exists in Communication : Human-machine ethics beyond the Actor-Network. *London School of Economics and Political Science,* (pp. 1–25). http://urn.kb.se/resolve?urn=urn:nbn:se:sh:diva-37239

Canas, K., Ubiera, B., Liu, X., &amp; Liu, Y. (2018). Scalable biomedical image synthesis with gan. *Proceedings of the Practice and Experience on Advanced Research Computing,* 1–3. https://doi.org/10.1145/3219104.3229261

Dafoe, A. (2015). On technological determinism. *Science, Technology, & Human Values, 40(6),* 1047–1076. https://doi.org/10.1177/0162243915579283

Hong, Y., Hwang, U., Yoo, J., &amp; Yoon, S. (2020). How generative adversarial networks and their variants work. *ACM Computing Surveys, 52(1),* 1–43. https://doi.org/10.1145/3301282

Luo, S. (2021). A survey on multimodal deep learning for image synthesis. *The 5th International Conference on Innovation in Artificial Intelligence,* 108–120. https://doi.org/10.1145/3461353.3461388

Kaartemo, V., & Helkkula, A. (2018). A systematic review of artificial intelligence and robots in value co-creation: Current status and future research avenues. *Journal of Creating Value, 4(2),* 211–228. https://doi.org/10.1177/2394964318805625

Kuleshov, A., Ignatiev, A., Abramova, A., & Marshalko, G. (2020). Addressing AI ethics through codification. *2020 International Conference Engineering Technologies and Computer Science (EnT).* 24-30. https://doi.org/10.1109/ent48576.2020.00011

Matthews, A. (2020). Blurring boundaries between humans and technology: Postdigital, postphenomenology and actor-network theory in Qualitative Research. *Qualitative Research in Sport, Exercise and Health, 13(1),* 26–40. https://doi.org/10.1080/2159676x.2020.1836508

O’Lemmon, M. (2020). The technological singularity as the emergence of a collective consciousness: An anthropological perspective. *Bulletin of Science, Technology & Society, 40*(1-2), 15–27. https://doi.org/10.1177/0270467620981000

Pinch, T. J., & Bijker, W. E. (1984). 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*(3), 399–441. http://www.jstor.org/stable/285355

Qiao, P., Dou, Y., Feng, W., Li, R., &amp; Chen, Y. (2017). Learning non-local image diffusion for image denoising. *Proceedings of the 25th ACM International Conference on Multimedia,* 1847–1855. https://doi.org/10.1145/3123266.3123370

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., & Norouzi, M. (2022). Palette: Image-to-image diffusion models. *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings,* 1–10. https://doi.org/10.1145/3528233.3530757

Vogel, K. M. (2021). Big Data, AI, platforms, and the future of the U.S. Intelligence Workforce: A research agenda. *IEEE Technology and Society Magazine, 40*(3), 84–92. https://doi.org/10.1109/mts.2021.3104384