DEFINING AND DESIGNING AN ETHICAL APPROACH TO GENERATIVE ARTIFICIAL INTELLIGENCE IN TEXT-TO-IMAGE MODELING

By

Drake A. Goodman (drakeg@wharton.upenn.edu)

An Undergraduate Thesis submitted in partial fulfillment of the requirements for the

JOSEPH WHARTON SCHOLARS

Faculty Advisor:

Amy J. Sepinwall (sepin@wharton.upenn.edu)

Associate Professor, Legal Studies & Business Ethics

THE WHARTON SCHOOL, UNIVERSITY OF PENNSYLVANIA

MARCH 2024

ABSTRACT

This paper addresses the ethical concerns that generative artificial intelligence (AI) in text-to-image modeling poses, specifically in protecting social equality and against discrimination. It first defines AI and explains the focus on generative AI. It then discusses the emergence of generative AI modeling, as well as prominent players in the generative AI text-to-image space. After explaining different ethical questions that have arisen in response to the rapid deployment of the technology, this paper establishes an ethical claim for why AI developers need to design these models to protect social equality and reduce stereotypes. This occurs in two ways. The first is explaining how the biases and stereotypes present in generative AI differ from the world before this technology existed. The second is establishing why amplifying stereotypes and biases is wrong, as well as why generative AI developers specifically have a moral obligation to protect social equality.

I. INTRODUCTION

Generative artificial intelligence (AI) has become an increasingly popular tool to produce images and videos, gaining traction especially in marketing and advertising. Some experts predict that AI-generated images will constitute 90 percent of all images available on the internet in the next few years (Nicolleti and Bass 2023). Additionally, by 2025, it is estimated that large companies will use generative AI tools to produce 30 percent of all marketing content and that AI could be used to create blockbuster movies using text-to-video prompts (Nicolleti and Bass 2023). It is apparent that generative AI will increasingly dominate the images and videos that consumers see. This is particularly concerning because these AI-generated images often depict skewed demographics, even as the images are becoming indistinguishable from real photos. In particular, these images lead to the proliferation of stereotypes about specific groups and identities, which can be defined as sets of beliefs about the characteristics, attributes, and behaviors of members of certain groups (Ashmore and Del Boca 2015). By amplifying stereotypes and marginalizing specific groups, AI-generated images have become an ethical concern.

This paper presents a novel examination of the emergence of text-to-image generative AI, the ethical concerns the technology poses, and the obligation that AI developers have to prevent discrimination and social inequality. Previous research addressing ethical concerns has been conducted around the use of AI in criminal justice, access to the financial system, healthcare, online content moderation, human resources, and education (Raso, Hilligoss, Krishnamurthy, Bavitz, and Kim 2018). In each of these cases, important decisions such as determining credit scores or diagnosing patients are directed by AI systems, with accompanying concerns around safe and effective systems, discrimination protections, data privacy, notice and explanation, and human alternatives (The White House Office of Science and Technology 2023).

The preliminary research into generative AI and the ethics surrounding this new technology outlines a myriad of ethical issues that broadly applies to generative AI: "(i) no regulation of the AI market and urgent need for regulation, (ii) poor quality, lack of quality control, disinformation, deep fake content, algorithmic bias, (iii) automation-spurred job losses, (iv) personal data violation, social surveillance, and privacy violation, (v) social manipulation, weakening ethics and goodwill, (vi) widening socio-economic inequalities, and (vii) AI technostress" (Wach et al. 2023). The existing research concerns itself with contexts where AI is used to make decisions, including what treatment a patient should receive, which loan applicant should receive a mortgage, etc. But AI is used not only as a decision-making tool but sometimes simply to generate new content (e.g., ChatGPT). Where there is no risk of fraud (e.g., cheating on course assignments), one might have thought that the ethical concerns of having AI simply create text or images were minimal. That is, digital images would seem much less consequential than digital decision-making in important life dimensions. This thesis suggests otherwise: the concerns about generative AI in the text-to-image space deserve serious consideration. This paper aims to draw out the problems and propose an ethical framework surrounding social equality to address them.

As we will see, generative AI threatens to amplify negative stereotypes. This paper will argue that there is an ethical obligation to protect social equality, by refraining from amplifying negative stereotypes in generative AI text-to-image modeling. This is especially important as there is increased scrutiny toward the industry from government entities such as the White House, which has recently proposed a "Blueprint for an AI Bill of Rights" and an executive order on the “Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence”. Included among its proposals are protections from algorithmic discrimination, which are of special relevance here. As the AI Bill of Rights outlines, people “should not face discrimination by algorithms and systems should be used and designed in an equitable way” which “occurs when automated systems contribute to unjustified different treatment or impacts disfavoring people based on their race, color, ethnicity [or] sex” (The White House Office of Science and Technology 2023). Such discrimination can even sometimes violate legal protections. Evidently, this is a priority for governments in ensuring a functioning and equitable society. The problem with the White House proposals is that they do not explicitly extend to generative AI. It first needs to be established that generative AI produces discrimination, which is one of the main aims of this paper. Secondly, the White House proposals do not directly extend to businesses because the government cannot regulate sexism and / or racism in advertising due to the First Amendment; thus, we need to rely on ethics to go beyond the law.

This paper will begin to explore these considerations, in Part II, by establishing a general definition of AI as well as understanding its implications as an emerging technology. Part III will discuss the emergence of generative AI text-to-image models, which will emphasize how these models work and the ethical issues that have arisen as a result of these rapid developments. After contextualizing these concepts, Part IV will establish an ethical framework for why generative AI should not amplify stereotypes and why developers have an obligation to prevent the algorithms from doing so. To do this, it will be important to understand how the world operated before the rise of generative AI to demonstrate the profound impact of generative AI and the true problems the technology poses. After establishing this connection, the paper will propose an ethical claim that the social inequality and stereotype amplification that generative AI in text-to-image modeling creates is morally wrong. It will do so by explaining how normalizing one demographic in positions of power impermissibly hurts women and people of color. It will argue further that AI developers have a moral obligation to protect social equality, which entails an obligation to ensure more representative AI images. The paper concludes in Part V by offering some practical recommendations.

II. GENERAL DEFINITION AND FOCUS ON GENERATIVE AI

General Definition of AI

It is important to define AI in order to lay out an ethical framework. The definition of AI has evolved over time as it has served different functions and goals as well as expanded its capabilities. Furthermore, developments in AI capabilities have changed how we think about the ever changing technology. One way to approach defining AI is to categorize it into different frameworks, with the most relevant ones being Function-AI and Principle-AI. Function-AI helps distinguish AI from other branches of computer science by associating it with cognitive functions identified in the human mind. In other words, the function that "maps a percept (input problem) into an action (output solution) in the computer is similar to that of a human" would be considered AI (Wang 2019, 12). One issue with this approach though is that it is too fragmented and compartmentalized rather than an integrated and cohesive system that can account for all components of AI simultaneously.

Another perspective is the Principle-AI framework, which defines intelligence as a "form of rationality that can make the best-possible decision in various situations, according to the experience or history of the system" (Wang 2019, 12). While similar to Function-AI in that this focuses on the function rather than actual percepts and actions, the input-output relationship in this framework is more general. In other words, this type of intelligence can use the agent's history of various situations to deal with a variety of problems rather than solve a single type of problem with its attached solution. One definition under the Principle-AI framework is that intelligence is a "strategy of problem-solving that is fundamentally different from computation" (Wang 2019, 20). Therefore, we can extract that AI is a machine or software that can solve a variety of problems fundamentally different from the method of computation. This problem-solving method is enabled by its experience or history in various situations. That said, some critics argue that this definition is too complicated and heterogeneous to obtain a clean definition of AI. Still, as the technology becomes more widespread and complex, this definition offers the most applicable and general understanding of AI.

Generally speaking, AI currently makes predicting abundant and inexpensive, which has allowed the technology to become more widespread in recent years. This trend will continue as AI expands to consumer usage (Agrawal, Gans, and Goldfarb 2017). There have been three technological developments that have enabled AI to solve business problems: advancement in algorithms, massive data, and increasing computational power and storage at low cost. In terms of its immediate impact, AI "is expected to increase the average annual revenues by more than ten trillion dollars all over the world and [is] expected to transform many industries" (Ergen 2019, 5). Due to the technology growing in adoption and becoming more integrated in people's everyday lives, the importance of ethical guardrails becomes clear as well.

III. EMERGENCE OF GENERATIVE AI

Within the broad landscape of AI, an increasingly popular form is generative AI. Generative AI is increasingly prevalent in the world, transforming how people work and interact.

In April 2023, 79 percent of employees (across all seniority levels) at corporations had at least some exposure to generative AI, and 22 percent of employees used it regularly at work (Chui, Yee, Hall, and Singla 2023). These figures are even higher when concentrating on the technology sector and North American region. This is not just a phenomenon at the individual level, but entire organizations support the adoption of generative AI as well. In fact, one-third of companies regularly use generative AI in at least one business function, which means that of companies that have begun adopting the technology, approximately 60 percent use it regularly. (Chui et al. 2023). Furthermore, the most common business function to adopt this technology within companies is marketing and sales.[1] While the most common use cases for generative AI in marketing are crafting the first drafts of text documents and personalized marketing, text-to-image marketing is emerging. It is predicted that by 2025, this technology will be able to produce deployable final drafts of image-based marketing content, and by 2030, these final drafts will be superior to the work of professional artists, designers, and photographers (J.P. Morgan Global Research 2023). It is clear that generative AI text-to-image models are transforming the marketing landscape and the general business world at a rapid pace.

Overview of Generative AI

Generative AI modeling is a technique that creates synthetic artifacts through a three step process: analyzing training examples, learning the key patterns and general distribution, and creating realistic outputs (Jovanović and Campbell 2022). For the first step, these models are pre-trained on a dataset, which means that they create results based on large amounts of data. In this process, the model will learn key patterns and trends, which will optimize its results. The

[1] 14 percent of companies use generative AI for marketing and sales, compared to 13 percent for product / service development, 10 percent for service operations, 4 percent for risk, 4 percent for strategy and corporate finance, 3 percent for human resources, 3 percent for supply chain management, and 2 percent for manufacturing.

data must be similar to the objective of the model; for example, if a model is designed to generate images from text, then the data it is trained on has to be text inputs that correlate to images. The more data that the model is pre-trained on, the more accurate and expansive it will be. When users enter an input into the model after this process, it will create realistic outputs based on those prompts.

While it has gained massive popularity and attention in recent years, generative AI is not a new phenomenon. Generative models in AI have been present since the 1950s with the development of Hidden Markov Models and Gaussian Mixture Models, which generated speech and time series (Cao et al. 2023). Generative models did not start seeing true success in performance until the advent of deep learning, which is training computers to process and analyze data as a human brain would do. Initially, deep generative models did not overlap much in terms of functionality; for example, a model that could generate short sentences could not necessarily generate long sentences (Cao et al. 2023). New techniques such as recurrent neural networks (RNNs), Long Short-Term Memory (LSTE), and Gated Recurrent Unit (GRU) helped mitigate this problem, which gave these models much more flexibility and functionality. Similarly, image generation algorithms historically relied on hand-designed features that had difficulty in creating diverse and complex images. Starting in 2014, the introduction of Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion generative models enabled algorithms to have more fine-grained control and generate higher-quality images (Cao et al. 2023). Eventually, in 2017, several of these approaches intersected through the introduction of the transformer architecture. This new intersection enabled multimodal models, which analyze and generate based on multiple modalities such as text and images (Cao et al. 2023). This transformer architecture truly revolutionized the AI generation landscape and enabled larger-scale training. Text-to-image and text-to-video generative AI models benefitted massively from this advancement, and models such as CLIP and DALL-E have been introduced since 2020. The speed of new developments and discoveries within the generative AI field has accelerated in recent years, which is why it is so important to understand and focus on.

Different Generative AI Models

There are multiple different generative AI text-to-image models, and several companies and individuals increasingly use them to assist in their work. These models range in style and capabilities, with some producing realistic images and others producing artistic images. The main players are DALL-E, Imagen, and Stable Diffusion.

DALL-E

DALL-E is OpenAI's text-to-image platform, which generates synthetic images based on users' input texts (Marcus, Davis, and Aaronson 2022). DALL-E 1 was initially launched in 2021, with DALL-E 2 launching in 2022 and DALL-E 3 launching in 2023 (O'Meara and Murphy 2023). The program's name is a combination of the surrealist Salvador Dali's name and Pixar's 2008 animated film *WALL-E* (which portrays a sentient robot who manages the garbage on a depopulated Earth in the future), which essentially merges the concepts of surrealism and machine learning / posthumanism (O'Meara and Murphy 2023). Each iteration of DALL-E has improved on the accuracy and diversity of content generation, which has mainly come from the data the model is trained on (Betker et al. 2023). OpenAI developers built DALL-E 3 based on the notion of caption improvement, which means that they trained the model on highly descriptive generated captions. This improvement in training data has allowed OpenAI to improve DALL-E's functionality and capabilities drastically over the past few years. In September 2022, DALL-E had 1.5 million users generating over 2 million images per day (OpenAI 2022), showing the widespread adoption of the program.

The program itself specializes in producing high quality images that have an artistic flair based on user text prompts. Overall, it produces images that suit the needs for amateurs with less strict expectations and parameters, while it still may not exactly fit the criteria that commercial professionals desire for their needs or clients. It can produce a diversity of artistic ranges, including light-hearted cartoons, peaceful impressionist paintings, and naturalistic everyday images (Marcus et al. 2022). While realistic styles are possible, the norm are non-realistic styles and surrealistic depictions. DALL-E allows users to depict creative situations in a variety of styles, which is deployed in both personal and commercial settings.

There are limitations with the model though, with one being safety and bias mitigations. DALLE-3 disproportionately represents individuals who appear White, female, and youthful (OpenAI 2023). While there have been improvements from previous iterations of DALL-E where the newest version now portrays individuals "in a more diverse manner that reflects a broad range of identities and experiences" (OpenAI 2023), stereotypes and biases still persist.

Imagen

Imagen is Google's text-to-image diffusion model that has a high degree of photorealism. The model was launched in 2022, and it is unique in its hyper realistic generated images. One key difference between Imagen and other text-to-image models is that instead of only using image-text data for model training, it also uses "text embeddings from [large language models

(LLMs)], pretrained on text-only corpora" (Saharia et al. 2022, 1). This means that complex text data is also effective for generating images from text prompts. Imagen 2 was launched in December 2023. Imagen is currently not widely available to the public, but it can be accessed through other companies such as Canva, which currently has over 170 million monthly users (Tirumalashetty 2023). The model therefore has several enterprise-ready capabilities that are being deployed and utilized, and it will continue to do so going forward.

Imagen 2 has several capabilities. Its main function is to produce high-quality, photorealistic, and high-resolution images based on natural language prompts. It can also generate these images based on prompts in multiple different languages, which adds to the complexity of the model (Tirumalashetty 2023). The main improvements between Imagen 2 and its predecessor are improved image-caption understanding, more realistic image generation, fluid style conditioning, and advanced inpainting and outpainting (Google DeepMind 2023). Further description was added to Imagen's training dataset, which helped the model learn different captioning styles and understand a wider array of possible user prompts. This has resulted in a better understanding of the relationship between words and images as well as added complexity and nuance in image generation. Imagen 2 also has more realistic image generation because the developers trained a specialized image aesthetics model based on "human preferences for qualities like good lighting, framing, exposure, sharpness, and more" (Google DeepMind 2023). By adding more weight in the model to these traits in the images in the training dataset, which means these images are valued higher in the decision making process, Imagen 2 has been able to produce higher quality images. The model also enables users to have more control and flexibility with the style of the generated images through adding a reference style in the text prompt. All of these functionalities make Imagen 2 enterprise-ready, especially in marketing and content creation.

That said, Imagen has similar downsides to DALL-E. This includes the ability for individuals to use the model for malicious purposes, including harassment and the spread of misinformation, which is why Google chose not to make Imagen 1 widely available to the public. Furthermore, one main limitation that the developers have acknowledged is the social bias present in the algorithm. Data audits have confirmed that the dataset that Imagen is trained on reflects social stereotypes, oppressive viewpoints, and derogatory associations to marginalized groups (Saharia et al. 2022). For example, one of the datasets that Imagen 1 trained on was LAION-400M, which contained racist slurs and harmful social stereotypes (Saharia et al. 2022). The model therefore amplifies these stereotypes and viewpoints in its image generation, which disproportionately impacts already marginalized individuals and communities. While some of these problems have been somewhat alleviated with Imagen 2, these biases and stereotypes still exist in the model.

Stable Diffusion

Stable Diffusion is another important player in the generative AI text-to-image modeling space. The model was launched by Stability AI in 2022, and the company launched Stable Diffusion 2.0 in November 2023. Overall, the Stable Diffusion model produces both surrealistic and photorealistic images, and its latest model can produce four HD images in less than three seconds. The technology has both personal and commercial applications, and it was even used to reproduce brain activity in medical research (Stability AI 2023). Stable Diffusion has garnered over ten million users, and its open access foundational model enables developers to download and build on top of the code for free (Stability AI 2023). Stability AI's goal for the model is to democratize the coding for generative AI text-to-image modeling, which is why it allows developers to access and build on its code without charge.

That said, there are still limitations with the model. The main one to highlight are the inherent biases present in the model, which a Bloomberg analysis highlights. This includes a disproportionate representation of specific demographics in certain roles and occupations, such as an overrepresentation of White male doctors and underrepresentation of female judges (Nicolleti and Bass 2023). These concerns are especially alarming considering how accessible the code is for other developers to access and build upon.

IV. ETHICAL QUESTIONS AMIDST THE EMERGENCE OF AI

After understanding the generative AI landscape and the problems that exist with stereotypes and potential social discrimination, it is important to analyze the ethical implications of this emerging technology. This section will explore the ethical concerns that generative AI poses as it relates to amplifying stereotypes and social discrimination. It starts with a general framework of ethical questions and considerations in the broad AI space. The next section will analyze the specific biases that generative AI has, further outlining the problem that exists. The following section will discuss what the world looked like before AI and the new ethical concerns that AI raises for stereotype amplification. After establishing these problems, this section will discuss the wrong of stereotyping and the moral obligation for developers to prevent further harms through generative AI.

General AI Ethics

AI systems are beginning to make autonomous decisions as humans do, and several questions arise from this phenomenon. One essential question surrounds the ethicality of AI decisions and outcomes. Within this question, there are three separate considerations: ethics by design, ethics in design, and ethics for design. Ethics by design involves “the technical / algorithmic integration of ethical reasoning capabilities as part of the behavior of an artificial autonomous system” (Dignum 2018, 2). Ethics in design involves “the regulatory and engineering methods that support the analysis and evaluation of the ethical implications of AI systems as these integrate or replace traditional social structures” (Dignum 2018, 2). Lastly, ethics for design involve “the codes of conduct, standards, and certification processes that ensure the integrity of developers and users as they research, design, construct, employ, and manage artificial intelligent systems” (Dignum 2018, 2). There are ethical questions at every stage of using AI, including pre-development and actual deployment of the technology, and each stage has tangible implications that impact various stakeholders. While legislative regulation is one solution to ensure the implementation of ethical guardrails in AI technology, self-regulation is more effective, practical, and sustainable (Buiten 2019). Consequently, there is more pressure on companies and developers to ensure ethical practices and deployment with the technology.

In terms of ensuring the AI acts ethically, Raso et al. (2018) identified three main areas where the problem arises: quality of training data, system design, and complex interactions. One reason that AI models have biases is because the data that it is trained on is biased, meaning there may be overrepresentation or underrepresentation of specific groups and scenarios. As a result, the AI model assumes this data is true and encompassing, and the results it will produce will reflect these inconsistencies. For the system design, developers can optimize certain variables and decide which variables the AI takes into consideration as it operates, which has tangible effects on the output the AI produces. Lastly, once an AI system is deployed, it might interact with an unfamiliar situation or environment, and this could lead to unforeseen outputs. As a result of this research, there have been heightened calls for developers to create guardrails and governments to increase regulation around AI.

While there are several ethical concerns regarding AI generally and generative AI in particular, one that has serious social consequences is the biases present in the algorithms and the social discrimination that can ensue. As AI becomes more integrated in decision making, biases and discrimination through the algorithms produce more severe consequences. For example, AI is being deployed to determine loan eligibility, diagnose patients, and facilitate portions of the criminal justice system (Ferrer, van Nuenen, Such, Coté, and Criado 2021). The AI is not always accurate though, and evidence shows that marginalized groups are disproportionately discriminated against in this process. This results in digital discrimination, where there is "unfair or unequal treatment of an individual (or group) based on certain protected characteristics (also known as protected attributes) such as income, education, gender, or ethnicity" (Ferrer et al. 2021, 72). This concern applies to generative AI in text-to-image modeling as well because it can portray certain demographics based on stereotypes, which has social implications and ramifications.

Generative AI and Bias

To further understand the bias present in these outputs, Bloomberg analyzed 5,100 images using Stable Diffusion, one of the leading AI text-to-image generator tools. The analysis found that when asked to create images based on job descriptions, "image sets generated for every high-paying job were dominated by subjects with lighter skin tones, while subjects with darker skin tones were more commonly generated by prompts like 'fast-food worker' and 'social worker'" (Nicolleti and Bass 2023). Additionally, most high-paying occupations were saturated with images depicting men, while low-paying jobs such as housekeepers and cashiers were primarily images of women. These depictions perpetuate biases: when compared to real-life statistics about who holds what kinds of jobs, the AI-generated images overrepresented males and those with lighter skin tones as holding higher paying jobs, while they overrepresented females and those with darker skin tones for lower paying jobs. Altogether, AI-generated images for higher paying professions such as teachers, architects, engineers, politicians, lawyers, CEOs, judges, and doctors vastly underrepresented women compared to reality, while women were overrepresented in lower paying jobs such as dishwashers, cashiers, housekeepers, and social workers (Nicolleti and Bass 2023). For example, while 34 percent of American judges are women, when prompted to create images of "judges", only 3 percent of the images produced by Stable Diffusion depicted women (Nicolleti and Bass 2023). The same holds true for AI representations of doctors: while women compose 39 percent of the profession, AI-generated results from Stable Diffusion created images where only 7 percent of the doctors were women (Nicolleti and Bass 2023). The amplification of stereotypes remains when depicting criminals and terrorists as well. In the Bloomberg analysis, more than 80 percent of images generated for the prompt "criminal" depicted people with darker skin tones despite people of color representing less than half of the prison population in the United States (Nicolleti and Bass 2023). Similarly, generated images of "terrorists" consistently portrayed Arabic men despite White men committing three times the number of terrorist acts in the United States compared to radical Islamists since September 11, 2001 (Nicolleti and Bass 2023). It is clear that the current technology is problematic because this will only further ostracize these marginalized groups and increase societal biases through the perpetuation of stereotypes.

These biases in the AI emerge due to the dataset that feeds inputs. In the case of Stable Diffusion, the company receives its raw data from LAION-5B, the world's largest openly accessible image-text dataset. The dataset consists of 5.85 billion CLIP-filtered image-text pairs, and its purpose is to allow LLMs to train on its data and democratize the general research on large-scale multi-modal models (Schuhmann et al. 2022). In fact, most of the generative AI text-to-image programs use LAION-5B as their training dataset, meaning the model tries to be as accurate as possible with its outputs based on data showing how a text prompt relates to an image. With more finetuning and data, these generated images become more accurate and realistic in depicting the inputted prompt. As these generated images become more indistinguishable from real images, ensuring that these images are truly representative of contemporary society and elevating marginalized groups will become essential.

The World Before AI

While it is clear that generative AI is changing the marketing and content-generation landscape, there is an important question to ask: how is this different from how everything worked before AI? While there are ethical questions surrounding the rise of generative AI, if it has improved social equality and decreased stereotypes compared to the previous norm, then perhaps it is a better alternative. Therefore, it is important to understand how people and roles were portrayed before the advent of generative AI, specifically in marketing and advertising. To understand this landscape before generative AI, we can analyze two scenarios: gender role portrayals in television advertisements and bias in search algorithms.

Gender Role Portrayals In Television Advertisements and Content

One important area to analyze is gender role portrayals. One demographic that is consistently misrepresented by generative AI text-to-image modeling are females in the workplace. Currently, a popular setting where consumers see gender role portrayals are in television advertisements and content. Karsay, Matthes, and Fröhlich (2020) conducted a study in Austria, and their findings suggest that if there is a female target group, television advertisements will contain the same or amplify stereotypes compared to a channel with a male target group. For example, women were more likely to be portrayed in domestic settings, appeared younger, and had less clothing compared to men. One difference compared to previous studies is that women and men were equally shown outdoors in the advertisements, which differed to 78 percent of male characters portrayed in the outdoors compared to 22 percent of females in 2008 (Karsay et al. 2020). Furthermore, the target audience amplified stereotypes in these advertisements. For example, advertisements intended for females generally portrayed younger women and older men, and they also showed women in more suggestive clothing than in the male channels (Karsay et al. 2020). Overall, this study shows that while there have been some improvements in decreasing gender stereotypes in television advertising over the years, there are some that still persist.

In the United States, these gender stereotypes persist as well. For example, females in advertisements and television content focus more on their appearance, are judged more often for their appearance, and are more likely to be sexualized (Ward and Grower 2020). In terms of the portrayal of personality, male characters are more likely than female characters to be physically aggressive and order others around, while female characters are more likely fearful, polite, frail, or romantic. Expanding on this notion, female characters are typically more family-oriented than their male counterparts (Ward and Grower 2020). In terms of occupation, males are more often portrayed in working roles while females are more often portrayed in domestic roles; even when portrayed in working roles, females are significantly underrepresented in STEM fields. Similarly, men are typically portrayed as incompetent in domestic roles (Ward and Grower 2020). As this meta analysis shows, gender portrayals in television advertising and content amplify existing stereotypes across several domains, including appearance, personality, and occupation. This is even true for advertisements and content targeted toward the youth, which influences them and their perceptions from a young age. That said, the portrayals of stereotypes has been decreasing over time as advocates push for better representation and consumers react more negatively to blatant stereotypes in advertising and television content. In general, there has been a decrease in gender-based stereotypes in commercials. There has also been a significant increase in gender parity with females receiving 55 percent screen time for youth television programs (Ward and Grower 2020). It is apparent that while gender-based stereotypes have historically been prevalent in television advertising and content, these trends have been decreasing over time.

Bias in Search Algorithms

Another focus area are search algorithms and search engines such as Google. While search engines are supposed to be neutral, they have historically amplified stereotypes through data discrimination. As Noble (2018) discusses, the "the combination of private interests in promoting certain sites, along with the monopoly status of a relatively small number of Internet search engines, leads to a biased set of search algorithms that privilege whiteness and discriminate against people of color, specifically women of color." One clear example of this issue is when searching up “Black girls” compared to “White girls”, the results for the former are sexually explicit terms or discussion forums of why that demographic is “sassy” or “angry”. The results for the latter are radically different. Through search engine results, these stereotypes are perpetuated to all users, affecting everyday lives and viewpoints.

It is worth noting though that while these search engine algorithms have biases, they do not seem to exhibit this bias toward specific stances. Typically, there is a liberal-leaning bias in ideology as well (Gezici, Lipani, and Yilmaz 2021). In other words, the biases that these search engine algorithms contain are complex, and they range from amplifying stereotypes about marginalized groups to containing a liberal ideology preference. There have been solutions to address the bias in search engines though, such as the development of meta-search engines. Meta engines aim to reduce bias by aggregating the results of traditional search engines and ignoring outliers to produce a “majority judgment” or “consensus” result (Maillé, Maudet, Simon, and Tuffin 2022). In other words, these search engines combine and analyze the results of several search engines to reduce bias and externalities. By employing meta engines, we can mitigate the biases that individual search engines potentially produce.

Evidently, marketing and advertising have historically amplified stereotypes in the context of genders and ethnicities. This has permeated multiple different mediums, ranging from television advertising to search engine results. That said, as people have become more conscientious of these issues, there have been tangible actions to mitigate the perpetuation of gender and ethnic stereotypes. This mitigation has not transpired in generative AI marketing and advertising, which is why it is so important to analyze.

What Makes the AI Case Different?

By realizing how the world worked before and how portrayals of diverse groups have been improving, we can better understand the harm generative AI in text-to-image modeling is causing and the ethical problems it poses. The initial component to this is understanding how stereotypes are ingrained in the algorithms. Ferrer et al. (2021) describe how the developers of AI models have biases and reproduce them in automated systems. This is because the models are trained on datasets that often reflect human judgments, priorities, and conceptual categories. Furthermore, the majority of developers do not extensively consider marginalized groups in the datasets or the coding of the AI modeling itself, so these groups will face biases in generative AI modeling (Ferrer et al. 2021). While the AI itself is not inherently a discriminatory entity, it is trained to be accurate based on the data it is presented with. If the data is biased, the AI model will be biased as well. With AI developing increasing prominence in content creation available to the public, these social inequalities become perpetuated and create feedback loops that amplify existing negative patterns in society (Mehrabi, Morstatter, Saxena, Lerman, and Galstyan 2021). As inequalities are encoded and replicated in these algorithms, there is an increased risk that legally protected groups are discriminated against. It is clear that there are clear risks within generative AI text-to-image modeling with regard to amplifying stereotypes and codifying social inequalities. Without any guardrails in place to mitigate this issue, public consumption of this content will lead to societal consequences.

The Wrong of Stereotyping

Given this context and background, it is important to understand the moral call to action that we are currently facing. This section will outline why actively *not* protecting minorities and marginalized groups in generative AI text-to-image modeling is morally wrong. For example, developing and perpetuating stereotypes in the workplace not only inhibits the growth and advancement of those targeted by these stereotypes, but it is also harmful to company performance. To reiterate, we can understand stereotypes as sets of beliefs about the characteristics, attributes, and behaviors of members of certain groups. Furthermore, the lack of representation of people with shared identities leads to internalized feelings of inferiority and discrimination. Evidently, as this section will discuss, not having guardrails to protect underrepresented groups has negative effects on those groups. In the AI case, these stereotypes are reinforced and perpetuated, especially because the decision making and human ownership is removed from the process. This section will further establish that due to the societal harm that generative AI text-to-image modeling can cause, AI developers have a moral obligation to mitigate and reduce this risk.

Why a World that Normalizes One Demographic in Positions of Power Impermissibly Injures Women and People of Color

The Negative Effects of a Lack of Diversity in the Workplace. There are two clear examples of how normalizing one demographic in positions of power impermissibly injures women and people of color. The first is the negative effects of having a lack of diversity in the workplace on general firm performance. An analysis of 366 companies found that companies in the bottom quartile for gender and ethnicity or race representation were 25 percent *less likely* to achieve superior financial performance than average companies (Hunt, Layton, and Prince 3). There are several reasons for this disparity, but the main drivers are that firms with less diversity have disadvantages in recruiting the best talent, decreased employee satisfaction, and worse decision making (Hunt et al. 9).

In terms of talent recruitment, there is scarce top talent, scarce talent for jobs that require significant human interaction, scarce emerging market talent, and geographic mismatches in talent demand and supply (Hunt et al. 10). Firms with less diverse teams that do not actively seek qualified diverse talent lose out on a significant talent pool that is becoming more prominent in advanced economies such as the United States. Furthermore, if diversity is low, there are also low levels of engagement in the workplace, which means that even when firms recruit talent, they have a difficult time maximizing their capabilities. These lower levels of engagement are more prominent for Millennials and Generation Z employees, who care more about a strong commitment to diversity. By not actively stressing an environment of inclusion for people of different backgrounds, firms can lose employee interest and engagement, further endangering general productivity (Hunt et al. 10).

A similar case is present for the employee satisfaction of women and historically marginalized minorities. When there is not a sufficient number of minority-group members at a firm, the confidence and self-esteem of those individuals can suffer (Hunt et al. 11). This is because there is not enough of a support group to foster positive attitudes and behaviors; additionally, prejudices can still persist. Furthermore, when diversity recruitment is viewed as a token effort, the psychological outcomes for the targeted groups are poorer (Hunt et al. 11). This can ostracize these minority groups and lead to persistent internalized feelings of inferiority.

Less diversity in the workplace also can lead to worse decision making due to a variety of factors. Generally speaking, a diversity of perspectives and thoughts leads to a more informed final decision that considers a more comprehensive list of risks and benefits (Hunt et al. 12). As such, more homogenous teams tend to make worse decisions. Another element to this phenomenon is firms that do not foster innate and acquired diversity stifle innovation. Innate diversity represents the idea that firms who have a workforce that is more representative of its customer base are better positioned to understand its needs and propose strong innovations. Acquired diversity develops the notion that leaders need to actively encourage unorthodox views and creative solutions through a culture of inclusivity and psychological safety. When firms do not have innate and acquired diversity, its employees are less likely to offer creative solutions or feel valued in the decision making process, which ultimately negatively impacts execution.

While there is a compelling case that less diverse firms are fundamentally worse positioned, it is not solely more diverse representation that leads to a better and more productive workplace. In certain circumstances, teams that are more homogenous produce lower-quality work, worse decision-making, lower team satisfaction, and less equality (Ely and Thomas 2020, 118). That said, there needs to be certain conditions met to foster better performance on more diverse teams: "when team members are able to reflect on and discuss team functioning; when status differences among ethnic groups are minimized; when people from both high- and low-status identity groups believe the team supports learning; and… when teams orient members to learn from their differences rather than marginalize or deny them" (Ely and Thomas 2020, 118). In other words, firms need to foster an environment where minorities are not only included, but also treated equally and enabled to both learn and teach others. Ely and Thomas propose a "learning-and-effectiveness" approach to maximize the benefits of diversity. This paradigm encourages firms to build trust, actively work against discrimination and subordination, embrace a wide range of styles and voices, and make cultural differences a resource for learning (Ely and Thomas 2020, 118-120). Overall, firms that have diverse teams but do not actively promote egalitarianism across backgrounds and identities to support women and historically marginalized ethnic groups will perform worse. Inequality limits companies' capacity for innovation and continuous improvement and suppresses the leadership potential of its employees. In this sense, there are two key findings in this research. The first is that less diverse workplaces limit the potential of its minority group members while also not reaching its full potential as a firm. The second is that with diverse teams but without a learning-and-effectiveness approach, firms still suffer these same problems because its diverse members do not feel included or heard. As a result, minority groups need to be included both physically and psychologically to truly unlock their potential and the firm's potential.

The Internalized Socially Inferior Perception Due to a Lack of Representation. A lack of diversity is not only bad for business; it also injures underrepresented employees. Overall, lack of adequate representation has members of the underrepresented groups internalize the belief that they are socially inferior and do not deserve to occupy respectable social roles. One theory that encapsulates this idea is the minority empowerment thesis, which suggests that minority representation strengthens representational links (Banducci, Donovan, and Karp 2004). On one hand, individuals benefit from seeing representation of a part of their identity, which could lead to increased performance, more active participation, and higher general confidence. There are also the negative consequences of the lack of representation. When groups with a shared identity see themselves not represented or represented negatively, there is a tendency to have less connection or motivation to favorable societal positions, or there is a tendency to have negative internal perceptions.

One can see examples of this phenomenon even in early childhood, as demonstrated in the famous doll study. This was initially conducted by Kenneth B. Clark and Mamie P. Clark in 1947. Their study showed that Black children internalized negative societal views toward themselves. The study occurred during segregation in the United States and had Black children from Northern and Southern states between the ages of three and seven answer eight questions relating to four dolls (two Black and two White). While approximately two-thirds of the children preferred to play with the light-skinned dolls over the darker-skinned dolls, the Black doll was chosen more frequently for negative qualities and characteristics (Byrd et al. 2017). Not only did this study show that children are cognizant of racial attitudes and have begun developing their own, but it showed that Black children more often associated negative traits with the Black dolls. This study also contributed to the Black self-hatred thesis. This theory posited that because of the racial prejudice and discrimination Black people have historically experienced in the United States, they internalize these societal negative attitudes, which could lead to lower self esteem and other negative consequences. There have since been numerous studies debating the validity of the self-hatred thesis (Baldwin, Brown, and Hopkins 1991; Banks 1976; Brand, Ruiz and Padilla, 1974; Rosenberg and Simmons 1971), but its legacy persists.

Byrd et al. replicated the doll study in 2017, several decades after desegregation and after racial attitudes had changed in the United States. They sampled 50 students between the ages of five and ten, and 47 of the students were Black, two were Latino, and one was White. They also had four dolls in the study: "Doll A was a White doll with dark hair and dark eyes with a non-Eurocentric appearance; b) Doll B was a White doll with blonde hair and blue eyes representing the Eurocentric appearance and in this case adapted by the American culture for standards of beauty; c) Doll C was a light-skin tone Black or biracial (i.e., African American and

White) with dark hair and dark eyes; and d) Doll D was a dark-skin tone African American with dark hair and dark eyes" (Byrd et al. 2017, 193). The authors asked ten questions, ranging from identifying dolls to associating dolls with preferences and attitudes. For the questions pertaining to positive attributes, the majority of participants chose the dolls that looked more similar to them. This shows a general shift since pre-segregation of children having a more positive self-concept. One hypothesis the authors propose is that more diverse dolls are displayed in stores today, which normalizes the acceptance of a variety of dolls (Byrd et al. 2017, 199). That said, when asked which doll is mean, the majority chose the dark-skinned doll. Additionally, when asked which doll is ugly, the participants chose the dark-skinned doll the most often. Overall, there is evidence that while attitudes have improved over the past several decades, there are still internalized negative attitudes due to societal influences. That said, these findings suggest that this can be improved with more diverse representation and the normalization of diversity in society.

While the doll studies have been helpful in understanding children's internalized attitudes, there are other cases where this phenomenon occurs. There are several examples of how this applies to leadership positions and occupations that marginalized groups have historically not participated in. In an undergraduate context, engineering is a field dominated by men and White and Asian students. Generally speaking, women and Black students encounter threats to their fit in such programs due to biased stereotyping and differential treatment (such as group work exclusion), which impacts their feeling of belonging. This has a negative impact on these groups declaring engineering as a field of study and completing it (Campbell-Montalvo et al. 2022). Another example of this is in educational leadership. When surveying underrepresented female educational leaders, Brown (2023) found that participants felt that their leadership practice was constrained by race and gender. In other words, they viewed that in order to succeed, they needed to go above and beyond what their counterparts are expected to do in order to feel accepted and command respect. This has partly led to the reason for why underrepresented females are severely underrepresented in educational leadership positions.

It is clear that even from a young age, people internalize perceptions based on who is portrayed occupying which roles. In the doll studies, children associated the different colored dolls with different traits, including the darker skinned doll with more negative traits. While these attitudes have improved since the initial doll study, many of these internalized feelings still persist. Furthermore, these negative internalized self-perceptions exist beyond childhood. When an individual is trying to pursue a field or position that is dominated by a group of a shared identity that makes the individual not feel a sense of belonging, there are significant obstacles including internalized feelings of not belonging or needing to prove oneself. It stands to reason that underrepresentation based on a part of someone's identity (such as ethnicity or gender) leads to a sense of not belonging and negative self perceptions. Not only can these internalized perceptions be exhausting and grueling, but they can lead to worse performance or the decision to not pursue a field or career.

What is the Harm in Not Having More Diversity?

The General Case

It is now clear that stereotyping is morally wrong, but what happens when there are not clear protections to ensure diversity in different settings? There is significant evidence showing the detriment a lack of protections around diversity can cause. In a meta-analytical review of negative stereotypes and devaluing content in the media, Appel and Weber (2021) found that members associated with negatively portrayed groups were also negatively affected. This is a result of the social identity threat, where individuals fear that they will conform to a negative stereotype through their own behavior. This threat leads to negative emotions and characteristics, physiological stress, and aversive thoughts and feelings. After analyzing several studies observing the impact of stereotypes portrayed in the media on groups associated with those identities, Appel and Weber (2021) found that this led to worse cognitive performance and less association with this part of their identity. In other words, stereotypes that are perpetuated by the media inherently harm individuals who are associated with the portrayed groups. To make matters worse, when traits are ascribed to members of a social category or group, this can disqualify them from roles whose requirements do not align with their existing stereotypes (Eagly and Koenig 2021). This is because stereotypes form from shared observations, which are strengthened from their apparent consensuality; furthermore, people often are biased to confirm their expectations and seek stereotype-consistent information (Eagly and Koenig 2021). As a result, if someone is applying for or seeking a role that is not consistent with societal stereotypes about that person, then he or she will have a much more difficult time overcoming those biases and preconceptions. The opposite is true too. When the roles shift for a particular group that do not align with the traits of an existing stereotype, the intensity of this stereotype decreases. For example, the implicit and explicit stereotypes associating women with family and the arts and men with careers and science have decreased over time as these roles have shifted (Eagly and Koenig 2021). That said, stereotypes still need to decrease for this transition to happen, or there will be severe resistance for individuals to obtain roles not consistent with the traits of their existing stereotypes.

These perpetuated stereotypes not only cause mental and physiological harm to individuals who are associated with the groups being stereotyped, but they directly impact employment opportunities as well. Under the theory of statistical discrimination, when employers do not have perfect information about the future productivity of job candidates, this incentivizes them to use easily observable ascriptive traits such as race or gender to infer their future productivity (Tilcsik 2021). This is not usually out of maliciousness; instead, profit-maximizing employers use all available information to make decisions, and when individual-specific information is limited, they defer to group membership as a proxy (Tilcsik 2021). When analyzing the impacts statistical discrimination has in labor markets through a simulation, Tilcsik (2021) discovered that the general idea that awareness of statistical discrimination strengthens employers' belief in the accuracy of stereotypes, acceptance of stereotyping, and rate of stereotyping. It is clear that in environments where decision making relies on limited information, the reliance on stereotypes can harm individuals who are associated with negative stereotypes. Furthermore, as stereotyping becomes justified or normalized, this practice can increase significantly as the merit of the individual becomes ignored more often.

These phenomena are directly observed when understanding how stereotypes can harm women. In general, the impact of gender-stereotyping extends to women across different levels of positions in male-dominated industries, but it is most pronounced in managerial and leadership roles (Tabassum and Nayak 2021). Overall, people identify successful leaders with traits that are more associated with men, such as leadership ability, competitiveness, self-confidence, objectivity, aggressiveness, forcefulness, ambition and desire for responsibility (Tabassum and Nayak 2021). Conversely, women are more associated with the concern and sympathetic treatment for others such as being affectionate, helpful, friendly, kind, sympathetic, interpersonally sensitive, gentle, and soft-spoken (Tabassum and Nayak 2021). This view has been consistent over time: in studies conducted since 1989, men and women have viewed men as more suitable for leadership positions; however, when management is less viewed as the domain of men, this perception decreases. As a result, it is clear how stereotypes reduce adequate opportunities to succeed in male-dominated industries, particularly in management roles, even when women are already working there. The intensity of this dynamic can decrease when stereotypes lessen and more women are normalized in these roles.

Overall, it is clear that the lack of guardrails in protecting against the propagation of stereotypes harms the groups associated with negative stereotypes. Stereotypes can lead to negative mental and physical effects that can inhibit individuals who feel targeted. They also can restrict access to occupations or advancement in the workplace. These observations are true for a variety of individuals and existing stereotypes, which can occur based on gender, race, sexuality, occupation, cultural identity, etc. Without having protections against the use and practice of these stereotypes, discrimination and negative self identity can occur and disproportionately affect historically marginalized groups who typically are associated with stereotypes that hinder them from social advancement.

The AI Case

While the above outlines the harm of not having protections around diversity, there are also clear implications of the harm of lacking adequate diverse protections in the generative AI space. If we understand stereotypes as a set of beliefs about the characteristics, attributes, and behaviors of members of certain groups, we can better analyze how generative AI amplifies them. For text-to-image generative AI, even when prompts are seemingly ordinary by simply mentioning traits, descriptors, occupations, or objects, they produce stereotypes about who fits in what role (Bianchi et al. 2023). Even when prompts avoid mentioning identity or demographic language, the outputs still produce stereotypes. In prompts that look for a trait or social role, their corresponding outputs tend to portray “Whiteness” as an ideal. As the Bloomberg analysis found, prompts that ask for occupations tend to amplify racial and gender stereotypes. Even when users ask for images with specific counter-stereotype language or institutions deploy system guardrails, these mitigation strategies are futile as stereotypes still persist (Bianchi et al. 2023). This all underscores how the generative AI text-to-image models directly create harmful stereotypes in even the most basic prompts and when prompts try to avoid producing stereotypes.

Another perspective is considering the macro-level consequences that social inequality perpetuated through generative AI can cause. In the United States, generative AI has entrenched social division and exacerbated social inequality, especially for historically marginalized groups. This is even more severe on a global scale: in countries such as Mozambique where approximately 90 percent of the population does not have internet access, they are not factored into the data for the AI modeling as much nor are they considered as an important target audience (Hagerty and Rubinov 2019). This underrepresentation in the data that the AI models are trained on excludes the needs and considerations of these individuals, which can further marginalize them as AI becomes more integrated in society and creates content and enables decision making. The amplification of social inequalities can increase general social instability, which threatens entire societies (Hagerty and Rubinov 2019). As such, individuals and organizations should be working to limit social inequalities to ensure societal stability.

Another perspective to consider is how generative AI interacts with human rights. Generative AI in text-to-image modeling can interact with specific rights, such as the freedom from discrimination and freedom of information. AI systems can inadvertently discriminate against marginalized groups in novel ways, and people may conceal their actions and information about themselves due to the fear of having this data be maliciously used in AI settings (Raso et al. 2018). Through the quality of training data, system design, and complex interactions, AI often can violate basic human rights that the law protects human actors from violating. Even while the technology can benefit the rights of certain groups, this often comes at the expense of marginalized groups (Raso et al. 2018). Furthermore, a single AI application can affect a variety of civil, political, economic, social, and cultural rights. What makes the case of AI more complicated is that while humans have the agency and free will to change their morals at any time, AI models do not have this luxury; instead, they require constant attention from developers to monitor and control the decision making of the system (Raso et al. 2018). Thus, if developers do not actively mitigate the biases and inequalities present in generative AI text-to-image modeling, there is a serious risk of violating human rights.

A final consideration is how this technology impacts individuals at a more micro-level. As Lloyd (2018, 2) outlines, “automation bias describes situations in which the AI fails to take social or cultural factors into consideration,” which is problematic because of the scale and speed that this bias can permeate. This is because when AI produces content or makes decisions, it leads to much less accountability than if a human actor had conducted these same actions. Consequently, the AI may negatively impact thousands or millions before the problem is mitigated. It is apparent that the consequences can be severe and far reaching if protections are not in place.

What is the Moral Obligation for AI Developers to Protect Social Equality in AI-Generated Images?

It is now clear that a world that normalizes one demographic in positions of power impermissibly injures women and people of color. It is also apparent that not having effective guardrails to decrease stereotypes harm these groups and their prospects for breaking down existing barriers. This is particularly true in the case of generative AI. The next important question is who has a moral responsibility to protect social equality in AI-generated images? This section will make the case that AI developers bear most of this responsibility.

Relational Ethics

The first step is establishing the moral obligation through relational ethics. Moral relationism is the idea that moral status is derived from an interactive property that warrants being realized or prized between one entity and another (Metz and Miller 2016). Rather than establishing the view that people perceive their environment in a passive way as rationalists would suggest, moral relationists assert that people "actively engage with the world around them in a meaningful and unpredictable way" (Birhane 2021, 5). Therefore, people have an active hand in the environment around them and impact others, which relates to developers designing AI that interacts with others. Under relational ethics, one would assess the morality of AI based on how well they "help us realize our role-based moral obligations prescribed by our social relations with others" (Zhu 2023, 61). This would establish that the morality of the AI decision-making reflects that of the developer, which prescribes an obligation for the developer to uphold. In taking this stance, we can establish a relational ethics framework that ensures developers do not perpetuate social inequalities and amplify stereotypes, which is based on the earlier claims of the harm of what happens when this does not occur.

AI Developers' Responsibility for the Harms Caused by their Generative AI Models

Based on this framework, there is an argument for why AI developers in text-to-image modeling need to bear responsibility in protecting social equality. Martin (2019) argues that algorithms are designed to perform tasks with a particular moral delegation in mind; similarly, computer scientists perform this same set of moral delegation when creating an algorithm. Computer scientists and developers deliberately choose how to train and create an algorithm to execute actions and choices. This shifts the view that algorithms are neutral to the view that they are value-laden with the choices of the developers. When delegating a task to technology, the responsibility is not alleviated (Martin 2019). One metaphor to this situation is how physicists look for "missing mass" in the universe just as sociologists or ethicists look for missing responsibility in a system of technology or individuals (Latour 1992, 152 - 153). Ignoring these moral delegations and missing masses does not make them any less important or less impactful, because when no one claims responsibility for the actions of an algorithm, there is no accountability. When firms create algorithms that are designed to minimize the role of individuals, they create inscrutable systems, which are more difficult to explain and understand.

In the case of AI, the responsibility is taken away from individuals, but it still remains; when this happens, the responsibility inevitably is given to the developers who created the algorithm, as they actively create the algorithm and design it to behave as it does and minimize the role of individuals in decision making (Martin 2019). This is articulated clearly where "by willingly creating an algorithm that works in a value-laden and particular manner, firms voluntarily become a party to the decision system and take on the responsibility of the decision to include the harms created, principles violated, and rights diminished by the decision system. How much responsibility and for what acts depends on how the algorithm is designed. In fact, as is argued here, the more the algorithm is constructed as inscrutable and autonomous, the more accountability attributed to the algorithm and the firm that designed the algorithm" (Martin 2019, 844). This same logic applies to generative AI in text-to-image modeling. Developers have an active role in designing and training models to make autonomous decisions that diminish the role of humans. Furthermore, as the relational ethics framework establishes, they have an active role in designing a model that interacts with others. The responsibility therefore does not shift away and disappear, but developers bear the brunt of it based on their choices that ultimately produced a model that has a value-laden inscrutable nature directly influenced by the developers.

AI Developers' Obligation to Protect Social Equality and Prevent Discrimination

Thus, if developers bear responsibility for harms caused by their generative AI text-to-image models, and these harms are clearly outlined as morally impermissible, do they have an obligation to prevent these harms? One way to answer this question is understanding the concept of justice failure. This is understood in the context of market failures, which are when there is a suboptimal efficiency of actual markets. In practice, when this occurs, ethical business leaders do not exploit these inefficiencies so that the underlying problem is not further exacerbated, which threatens the very existence of the system. There is also an implied morality in the market, where firms fail not solely on the grounds of efficiency, but by the standards of values and principles we collectively agree on (Norman 2014, 27). Justice failure serves as a complement to market failure, and it refers to the failure to achieve a morally desirable form of equality. Similar to how it is actively discouraged for firms to profit off of market failures to avoid contradicting the intrinsic morality of the market, firms should not intentionally profit off of justice failures to avoid contradicting the implied morality of the larger social scheme that ultimately justifies the morality of the market. This rests on the view that society should foster equal opportunity for its citizens, yet historical factors have conspired to prevent this (Singer 2018). Companies inadvertently do this when using generative AI because monitoring every image or video they produce is too costly and inefficient. Unfortunately, in doing so, they profit off of justice failures. When confronted with social justice failures, businesses not only should follow the law in protecting social equality and against discrimination, but they should actively combat the inequality they see in order to prevent the exacerbation of more social justice failures (Singer 2018). AI developers have a similar obligation to confront justice failures rather than ignore or exacerbate them. They also operate as corporations that are profiting from their technology. Furthermore, they have direct control over the model they are producing and providing access to at scale. Failing to combat these justice failures will only exacerbate the failures that our society and markets need to minimize.

V. POTENTIAL SOLUTIONS

There are different approaches AI developers can take to ensure these moral obligations are met. From a design perspective, AI developers can implement a loss function to minimize the stereotypes that are being perpetuated by AI-generated images. This technique computes the distance between the expected results and the actual results of an AI model, and it tries to optimize the model by reducing the distance and minimizing the loss function. One way to minimize the disproportionate representation of a type of person in an occupation or specific position is to change the weighted probabilities of outputs. For example, Qian, Muaz, Zhang, and Hyun (2019) effectively reduced gender bias in AI text generation by equalizing the probabilities of predicting gendered word pairs such as “he” and “she”. There are other techniques that AI developers can use as well to mitigate these biases, such as data preprocessing, but the main conclusion is that changes in the model design are possible to create a more ethical model. Developers do have to be careful with this approach though, as they can overcorrect which can also cause problematic issues. For example, Google’s Gemini, which is its AI chatbot, overcompensated in February 2024 by producing historically inaccurate images by placing a diverse range of people in inaccurate settings. This included portraying a Black woman as an American founding father and Black and Asian people as Nazi-era German soldiers (O’Brien 2024). As this error shows, this process requires multiple iterations and fine-tuning to ensure an unbiased and accurate model that can be used en masse.

Related to this, another solution is creating a comprehensive framework to ensure AI development teams are equipped to mitigate these biases. This could come in many shapes and forms. One fundamental reason that these disparities and bias in the AI outputs occur in the first place is because there is a lack of gender and social diversity on AI development teams, which is compounded with the demands of producing this technology as fast as possible (Khensani and Twinomurinzi 2023). One consequence of this dynamic is that teams could conceal the biases present in the models as they rush to release the newest versions to appease stakeholders. Even pre-deployment, there are significant bias risks because the testing phase is rarely representative of the diverse societal groups that are often negatively affected by the outputs of the AI model (Xivuri and Twinomurinzi 2023). To mitigate these risks, AI firms can prioritize ethical considerations when designing their models, receive more input and feedback from stakeholders impacted by the AI model, and emphasize transparency and explainability in their models while ensuring adequate testing measures to reduce bias before deployment. This will allow AI developers to truly consider and prioritize mitigating the harms that marginalized groups face from the outputs of the AI models. Furthermore, it will ensure a more relational approach that will create a more comprehensive, effective, and responsible use of the AI models.

There can also be a conceptual shift. This shift would focus on relational ethics when designing the generative AI text-to-image models, placing emphasis on people who are marginalized and vulnerable, which makes them the most impacted by AI. By shifting to a relational approach, AI developers and companies can seek more input and feedback when designing these systems. This would give more power to the people affected by the AI, therefore giving them more autonomy and creating a more equitable system (Mhlambi and Tiribelli 2023). By shifting to this conceptual approach, AI developers would be able to proactively recognize and combat the negative stereotypes that their algorithms produce.

While these solutions approach the problem from different angles, they all have the potential to ensure AI developers are creating ethical AI models that do not perpetuate stereotypes and harm already marginalized groups. Further research should be conducted to evaluate these potential solutions in more depth in the context of generative AI in text-to-image modeling to measure their feasibility and effectiveness. Nevertheless, it is important to understand that generative AI text-to-image modeling amplifies stereotypes and threatens social equality. There are risks with this phenomenon as more users adopt this technology to assist in advertising and content generation that are even more impactful than how the world operated before. By actively mitigating the harm their models can cause, AI developers will fulfill their moral obligation to protect social equality and against discrimination.

BIBLIOGRAPHY

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb. “What to expect from artificial intelligence.” (2017): 1486511104226.

Appel, Markus, and Silvana Weber. “Do mass mediated stereotypes harm members of negatively stereotyped groups? A meta-analytical review on media-generated stereotype threat and stereotype lift.” *Communication Research* 48, no. 2 (2021): 151-179.

Arora, A., M. Barrett, E. Lee, E. Oborn, and K. Prince. “Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization.” *Information and Organization* 33, no. 3 (2023): 100478.

Ashmore, Richard D., and Frances K. Del Boca. “Conceptual approaches to stereotypes and stereotyping.” *In Cognitive processes in stereotyping and intergroup behavior*, pp. 1-35. Psychology Press, 2015.

Back, Lindsey Therese. “Empowerment of underrepresented racial/ethnic minority college students in the United States: Developing and testing the college student empowerment scales for racial/ethnic minorities.” PhD diss., DePaul University, 2014.

Baldwin, Joseph A., Raeford Brown, and Reginald Hopkins. “The black self-hatred paradigm revisited: An Africentric analysis.” (1991).

Banducci, Susan A., Todd Donovan, and Jeffrey A. Karp. “Minority representation, empowerment, and participation.” *The Journal of Politics* 66, no. 2 (2004): 534-556.

Banks, W. Curtis. “White preference in Blacks: A paradigm in search of a phenomenon.” *Psychological Bulletin* 83, no. 6 (1976): 1179.

Betker, James, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang et al. "Improving image generation with better captions." *Computer Science. https://cdn. openai.com/papers/dall-e-3. pdf* (2023).

Bianchi, Federico, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale." *In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1493-1504. 2023.

Birhane, Abeba. "Algorithmic injustice: a relational ethics approach." *Patterns* 2, no. 2 (2021).

Brand, Elaine S., Rene A. Ruiz, and Amado M. Padilla. "Ethnic identification and preference: A review." *Psychological Bulletin* 81, no. 11 (1974): 860.

Brown, Kinasha. "Herstory 101: Examining the Representation Gap of BIPOC Women Educational Leaders." PhD diss., Northern Illinois University, 2023.

Buiten, Miriam C. "Towards intelligent regulation of artificial intelligence." *European Journal of Risk Regulation* 10, no. 1 (2019): 41-59.

Byrd, Diane, Yasmin R. Ceacal, Jansen Felton, Carmen Nicholson, David Martin Lakendra Rhaney, Nakia McCray, and Jaceline Young. "A modern doll study: Self concept." *Race, Gender & Class* 24, no. 1-2 (2017): 186-202.

Campbell‐Montalvo, Rebecca, Gladis Kersaint, Chrystal AS Smith, Ellen Puccia, John Skvoretz, Hesborn Wao, Julie P. Martin, George MacDonald, and Reginald Lee. "How stereotypes and relationships influence women and underrepresented minority students' fit in engineering." *Journal of Research in Science Teaching* 59, no. 4 (2022): 656-692.

Cao, Yihan, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. "A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT." *arXiv preprint arXiv:2303.04226* (2023).

Checketts, Levi. "Artificial Intelligence and the Marginalization of the Poor." *Journal of Moral Theology* 11, no. Special Issue 1 (2022): 87-111.

Chui, Michael, Lareina Yee, Bryce Hall, and Alex Singla. "The state of AI in 2023: Generative AI's breakout year." (2023).

Clark, Kenneth B., and Mamie K. Clark. "Skin color as a factor in racial identification of Negro preschool children." The Journal of Social Psychology 11, no. 1 (1940): 159-169.

Crawford, Kate. *The Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

Dignum, Virginia. "Ethics in artificial intelligence: introduction to the special issue." *Ethics and Information Technology* 20, no. 1 (2018): 1-3.

Eagly, Alice H., and Anne M. Koenig. "The vicious cycle linking stereotypes and social roles." *Current Directions in Psychological Science* 30, no. 4 (2021): 343-350.

Ely, Robin J., and David A. Thomas. "Getting serious about diversity." *Harvard Business Review* 98, no. 6 (2020): 114-122.

Ergen, Mustafa. "What is artificial intelligence? Technical considerations and future perception." *Anatolian J. Cardiol* 22, no. 2 (2019): 5-7.

Ferrer, Xavier, Tom van Nuenen, Jose M. Such, Mark Coté, and Natalia Criado. "Bias and discrimination in AI: a cross-disciplinary perspective." *IEEE Technology and Society Magazine* 40, no. 2 (2021): 72-80.

Gezici, Gizem, Aldo Lipani, Yucel Saygin, and Emine Yilmaz. “Evaluation metrics for measuring bias in search engine results.” *Information Retrieval Journal* 24 (2021): 85-113.

Giovanola, Benedetta, and Simona Tiribelli. “Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms.” *AI & society* 38, no. 2 (2023): 549-563.

Google DeepMind. “Imagen 2: Our most advanced text-to-image technology.” Google DeepMind, 2023. https://deepmind.google/technologies/imagen-2/.

Hacker, Philipp. “Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law.” *Common Market Law Review* 55, no. 4 (2018).

Hagerty, Alexa, and Igor Rubinov. “Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence.” *arXiv preprint arXiv:1907.07892* (2019).

Hunt, Vivian, Dennis Layton, and Sara Prince. “Diversity matters.” *McKinsey & Company* 1, no. 1 (2015): 15-29.

Jovanovic, Mladan, and Mark Campbell. “Generative artificial intelligence: Trends and prospects.” *Computer* 55, no. 10 (2022): 107-112.

J.P. Morgan Global Research. “Is generative AI a game changer?” J.P. Morgan, March 20, 2023. https://www.jpmorgan.com/insights/global-research/artificial-intelligence/generative-ai#:~:text=Generative%20AI%20tools%20reduce%20the,new%20business%20models%20and%20applications.

Karsay, Kathrin, Jörg Matthes, and Valerie Fröhlich. “Gender role portrayals in television advertisements: Do channel characteristics matter?.” *Communications* 45, no. 1 (2020): 28-52.

Latour, Bruno. “Where are the missing masses? The sociology of a few mundane artifacts.” *Shaping technology/building society: Studies in sociotechnical change* 1 (1992): 225-258.

Lloyd, Kirsten. “Bias amplification in artificial intelligence systems.” *arXiv preprint arXiv:1809.07842* (2018).

Maillé, Patrick, Gwen Maudet, Mathieu Simon, and Bruno Tuffin. “Are search engines biased? Detecting and reducing bias using meta search engines.” *Electronic Commerce Research and Applications* (2022): 101132.

Marcus, Gary, Ernest Davis, and Scott Aaronson. “A very preliminary analysis of DALL-E 2.” *arXiv preprint arXiv:2204.13807* (2022).

Martin, Kirsten. “Ethical implications and accountability of algorithms.” *Journal of business ethics* 160 (2019): 835-850.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. “A survey on bias and fairness in machine learning.” *ACM Computing Surveys (CSUR)* 54, no. 6 (2021): 1-35.

Metz, Thaddeus, and Sarah Clark Miller. “Relational ethics.” *The international encyclopedia of ethics* (2016): 1-10.

Mhlambi, Sabelo, and Simona Tiribelli. “Decolonizing AI ethics: Relational autonomy as a means to counter AI Harms.” *Topoi* (2023): 1-14.

Minsky, Marvin L. “Introduction to the Comtex microfiche edition of the early MIT Artificial Intelligence Memos.” *AI Magazine* 4, no. 1 (1983): 19-19.

Nicoletti, Leonardo, and Dina Bass. “Humans Are Biased. Generative AI Is Even Worse.” Bloomberg.com, 2023. https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

Noble, Safiya Umoja. “Algorithms of Oppression.” In *Algorithms of Oppression*. New York University Press, 2018.

Norman, Wayne. “Is there ‘a point’ to markets? A response to Martin.” *Business Ethics Journal Review* 2, no. 4 (2014): 22-28.

O’Brien, Matt. “Google says its AI image-generator would sometimes ‘overcompensate’ for diversity.” APNews.com, 2024. https://apnews.com/article/google-gemini-ai-chatbot-imagegenerator-race-c7e14de837aa65dd84f6e7ed6cfc4f4b.

O’Meara, Jennifer, and Cáit Murphy. “Aberrant AI creations: co-creating surrealist body horror using the DALL-E Mini text-to-image generator.” *Convergence* 29, no. 4 (2023): 1070-1096.

O’Neil, Cathy, and Hanna Gunn. “Near-term artificial intelligence and the ethical matrix.” *Ethics of Artificial Intelligence* (2020): 235-69.

OpenAI. “Dall·E Now Available without Waitlist.” OpenAI, September 28, 2022. https://openai.com/blog/dall-e-now-available-without-waitlist.

OpenAI. “DALL·E 3 System Card.” *OpenAI* (2023).

Qian, Yusu, Urwa Muaz, Ben Zhang, and Jae Won Hyun. “Reducing gender bias in word-level language models with a gender-equalizing loss function.” *arXiv preprint arXiv:1905.12801* (2019).

Raso, Filippo A., Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Levin Kim. “Artificial intelligence & human rights: Opportunities & risks.” *Berkman Klein Center Research Publication* 2018-6 (2018).

Rosenberg, Morris, Simmons R. Black, and White Self-Esteem. “The urban school child.” *Washington, DC: American Sociological Association* (1971).

Saharia, Chitwan, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour et al. “Photorealistic text-to-image diffusion models with deep language understanding.” *Advances in Neural Information Processing Systems* 35 (2022): 36479-36494.

Schuhmann, Christoph, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes et al. “Laion-5b: An open large-scale dataset for training next generation image-text models.” *Advances in Neural Information Processing Systems* 35 (2022): 25278-25294.

Singer, Abraham. “Justice failure: Efficiency and equality in business ethics.” *Journal of Business Ethics* 149 (2018): 97-115.

Stability AI. “Celebrating one year(ish) of Stable Diffusion … and what a year it’s been!” Stability AI, October 3, 2023. https://stability.ai/news/celebrating-one-year-of-stable-diffusion.

Tabassum, Naznin, and Bhabani Shankar Nayak. “Gender stereotypes and their impact on women’s career progressions from a managerial perspective.” *IIM Kozhikode Society & Management Review* 10, no. 2 (2021): 192-208.

The White House Office of Science and Technology. “Blueprint for an AI Bill of Rights.” The White House, 2023. https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

Tilcsik, András. “Statistical discrimination and the rationalization of stereotypes.” *American Sociological Review* 86, no. 1 (2021): 93-122.

Tirumalashetty, Vishy. “Discover Imagen 2, our most advanced text-to-image technology.” Google Cloud, December 13, 2023. https://cloud.google.com/blog/products/ai-machine-learning/imagen-2-on-vertex-ai-is-now-generally-available.

Wach, Krzysztof, Cong Doanh Duong, Joanna Ejdys, Rūta Kazlauskaitė, Pawel Korzynski, Grzegorz Mazurek, Joanna Paliszkiewicz, and Ewa Ziemba. “The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT.” *Entrepreneurial Business and Economics Review* 11, no. 2 (2023): 7-30.

Wang, Pei. “On defining artificial intelligence.” *Journal of Artificial General Intelligence* 10, no. 2 (2019): 1-37.

Ward, L. Monique, and Petal Grower. “Media and the development of gender role stereotypes.” *Annual Review of Developmental Psychology* 2 (2020): 177-199.

Xivuri, Khensani, and Hosanna Twinomurinzi. “How AI developers can assure algorithmic fairness.” *Discover Artificial Intelligence* 3, no. 1 (2023): 27.

Zhu, Qin. “Just Hierarchy and the Ethics of Artificial Intelligence: Two Approaches to a Relational Ethic for Artificial Intelligence.” *Ethical Perspectives* 30, no. 1 (2023): 59-000.