THE COLUMBIA
SCIENCE & TECHNOLOGY
LAW REVIEW

VOLUME XXV STLR.ORG SPRING 2024

FAIRNESS & PRIVACY IN AN AGE OF GENERATIVE AI

Alice Xiang[*]

Generative AI technologies have made tremendous strides recently and have captured the public's imagination with their ability to mimic what was previously thought to be a fundamentally human capability: creativity. While such technologies hold great promise to augment human creativity and automate tedious processes, they also carry risks that stem from their development process. In particular, the reliance of foundation models on vast amounts of typically uncurated, often web-scraped training data has led to concerns around fairness and privacy. Algorithmic fairness in this context encompasses concerns around potential biases that can be learned by models due to skews in their training data and then reflected in their generated outputs. For example, without intervention, image generation models are more likely to generate images of lighter skin tone male individuals for professional occupations and images of darker skin tone female individuals for working class occupations. This further raises questions around whether there should be legal protections from such pernicious stereotypical representations. Privacy is also a concern as generative AI models can ingest large amounts of personal and biometric information in the training process, including face and body biometrics for image generation and voice biometrics for speech generation. This Essay will discuss the types of fairness and privacy concerns that generative AI raises and the existing landscape of legal protections under anti-discrimination law and privacy law to address these concerns. This Essay argues that the proliferation of generative AI raises challenging and novel questions around (i) what protections should be offered around the training data used to develop such systems and (ii) whether

[*] Global Head (VP) of AI Ethics, Sony Group Corporation, and Lead Research Scientist, Sony AI. Thanks to Tiffany Georgievski and Swami Sankaranarayanan for their comments and suggestions on earlier versions of this essay, and to Professor Talia Gillis and the attendees of the 2023 Columbia Science & Technology Law Review Symposium, "Accountability and Liability in Generative AI: Challenges and Perspectives," for their helpful feedback on the piece. Thanks also to the staff of the Columbia Science & Technology Law Review for their excellent editorial support.

representational harms should be protected against in an age of AI-generated content.

I. INTRODUCTION

Since the release of ChatGPT, generative AI has garnered tremendous attention from the media, public, policymakers, developers, and investors. The concept of "generative AI," however, is by no means new. Early generative AI models can be traced to Hidden Markov Models and Gaussian Mixture Models, which were developed in the 1950s.[1] In 2014, the emergence of Generative Adversarial Networks helped popularize the concept of generative AI in the AI research community, but the explosion in public interest did not come until late 2022.[2]

Concern about the ethics of these models is also not new. One of the earliest examples of viral AI ethics controversies was Microsoft's Tay—a chat bot that was released in 2016, only to be quickly pulled by the company after users taught the model to spew racist, highly inflammatory content.[3] Concerns about deepfake technologies have also been in the popular discourse for several years, with a convincing deepfake of Tom Cruise going viral in 2021 and raising prescient questions about mis- and dis-information and the impact of AI on actors and other creatives.[4] Even though debates about the potential harms of generative AI technologies have been ongoing within the tech community, the recent attention to these technologies has elevated these discussions in the public sphere, with growing existential dread about the potential for AI not only to automate mundane, repetitive work but also creative work that seemed fundamentally human.

Since this recent surge in public attention, there has been a growing chorus calling for regulation of generative AI technologies, with EU regulators scrambling to update the draft of the EU AI Act to include language targeting such

[1] Yihan Cao et al., *A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT*, 37.4 J. ACM 111, 111.4 (2018), https://perma.cc/NJ55-Y7KM.

[2] F. J. García-Peñalvo & A. Vázquez-Ingelmo, *What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI*, INT'L J. INTERACTIVE MULTIMEDIA & A.I. 7, 12 (2023), https://perma.cc/N8L4-WETQ.

[3] Peter Lee, *Learning from Tay's Introduction*, MICROSOFT: OFFICIAL MICROSOFT BLOG (Mar. 25, 2016), https://perma.cc/R3MF-XB9G; Elle Hunt, *Tay, Microsoft's AI Chatbot, Gets a Crash Course in Racism from Twitter*, THE GUARDIAN (Mar. 24, 2016), https://perma.cc/T3S2-6WL6.

[4] *See* Scott Stump, *Man Behind Viral Tom Cruise Deepfake Videos Calls the Technology 'Morally Neutral,'* TODAY (Dec. 28, 2021), https://perma.cc/UA6Z-5KK9.

technologies,[5] and Congress members proposing bills[6] and engaging with the founders of major tech companies on potential regulation.[7] Given the current climate of uncertainty around these new technologies and the desire to regulate them, it is important to consider what their realistic harms are, whether current laws and regulations might protect against these harms, and where there might be gaps in protections. This Essay will map out these issues for two major categories of AI harms: bias/discrimination and privacy. First, I will discuss the problems of bias in generative AI technologies, with a particular focus on allocative versus representational harms. Then I will discuss current anti-discrimination doctrine in the U.S., particularly the lack of direct protections against representational harms, and explore whether such harms can be protected against under current laws governing content moderation and speech. Second, I will discuss the privacy harms associated with the training data and outputs for generative AI models and the lingering uncertainties around how privacy law might be applied in this context.

II. BIAS IN GENERATIVE AI

Bias in AI and the accompanying anti-discrimination law considerations have been a topic of scholarship over the past several years.[8] As a preliminary question, it is worth considering what new issues, if any, generative AI technologies raise in the anti-discrimination context. Much of the existing algorithmic fairness scholarship focuses on human-centric discriminative or classification-based AI technologies. These are AI models that learn patterns from data with the goal of classifying individuals, often in the context of risk assessment. For example, much of the early explosion in interest in algorithmic bias stemmed from ProPublica's revelation that the COMPAS recidivism risk algorithm had a false positive rate that was twice as high for Black defendants as for White defendants.[9] In the context of deep learning, the seminal Gender Shades paper showed much higher rates of misclassification for darker-skin female individuals compared with lighter-skin or male individuals for major commercial gender classification models.[10]

[5] *See* Thibau Duquin, *EU Artificial Intelligence Act and Generative AI – An Update*, STIBBE (June 14, 2023), https://perma.cc/9P7N-F7RY; Marie Barani & Peter Van Dyck, *Generative AI & the EU AI Act - A Closer Look*, JD SUPRA (Aug. 25, 2023), https://perma.cc/J2E4-YPFC.

[6] *See, e.g.*, Chris Coons et al., *Nurture Originals, Foster Art & Keep Entertainment Safe (NO FAKES) Act*, https://perma.cc/AV74-WRPM.

[7] Matt Laslo, *The US Congress Has Trust Issues. Generative AI Is Making It Worse*, WIRED (Sep. 13, 2023), https://perma.cc/44WX-5N4B.

[8] *See, e.g.*, Alice Xiang, *Reconciling Legal and Technical Approaches to Algorithmic Bias*, 88 TENN. L. REV. 649 (2021), https://perma.cc/H3T9-NQC8; Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016), https://perma.cc/ADK2-RDGB; Jon Klenberg et al., *Discrimination in the Age of Algorithms*, 10 J. OF LEGAL ANALYSIS 113 (2018), https://perma.cc/Q57G-2BHQ.

[9] Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://perma.cc/V86C-Q9LC.

[10] Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY (2018), at 10-12 https://perma.cc/DJC5-JV7C.

The type of harm most strongly associated with discriminative AI technologies is allocative harm, which stems from individuals being allocated different opportunities, resources, or experiences based on a protected attribute. In 2018, Amazon scrapped its resume filtering algorithm after discovering that the model systematically penalized female applicants.[11] In addition, many of the highest profile examples of algorithmic bias have been in the criminal justice context, where higher misrecognition rates for minorities by facial recognition models have led to wrongful arrests.[12] In these high-risk contexts, differential performance of the AI model leads to a misallocation of important resources, such as job opportunities and freedom.

Allocative harms are also relevant in the generative AI context, albeit less directly. For example, if a text generation model is asked whether a candidate should be moved to the next round of interviews or whether an individual should be detained without bail while awaiting trial, the model could directly cause allocative harm similar to the examples above. What is distinct, however, about generative AI is that the goal is typically not simply to automate many yes-no decisions or classifications but rather to create richer content, such as text, images, speech, or videos.

Given the typical goal of generative AI models, most of the literature thus far about bias in generative AI models has focused not on allocative harms but instead on representational harms, or harms that result from inaccurate or stereotypical representations. For example, an analysis featured in Bloomberg illustrated how popular generative AI models tended to generate images of lighter-skin and male subjects for high-paying occupations like, "architect," "lawyer," and "CEO" as opposed to darker-skin and female subjects for lower-paying occupations like "fast-food worker" and "social worker."[13] The extent of this bias exceeded societal biases, with women comprising only 3% of the generated images of "judges" despite comprising 34% of U.S. judges.[14] Moreover, over 80% of images generated using the keyword "inmate" featured subjects with darker skin tones.[15] A popular AI Renaissance portrait generator made individuals in the images look more white, with lighter skin tones and higher nose bridges.[16] In addition, biases have been

[11] Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*, REUTERS (Oct. 10, 2018), https://perma.cc/WDM5-JABE.

[12] *See* Khari Johnson, *How Wrongful Arrests Derailed 3 Men's Lives*, WIRED (Mar. 7, 2022), https://perma.cc/ZUN6-44CU.

[13] Leonardo Nicoletti & Dina Bass, *Humans Are Biased. Generative AI Is Even Worse*, BLOOMBERG (2023), https://perma.cc/H9VX-84LK.

[14] *Id.*

[15] *Id.*

[16] Morgan Sung, *The AI Renaissance Portrait Generator Isn't Great at Painting People of Color*, MASHABLE (July 23, 2019), https://perma.cc/UHD7-QFD4; Edward Ongweso Jr., *Racial Bias in AI Isn't Getting Better and Neither Are Researchers' Excuses*, VICE (July 29, 2019), https://perma.cc/N7UK-TJ8A.

found in how generative AI models can provide a very limited and stereotypical representation of the styles of famous artists like Van Gogh and Cezanne.[17]

These representational harms are worrisome given the explosion of AI-generated content, with some experts predicting that as much as 90% of online content might be synthetically generated by 2026.[18] To the extent that the world as we know it will increasingly be portrayed by AI models, it is worth questioning the acceptability of content that further entrenches and amplifies societal stereotypes. How will these representations affect people's views of society, people's beliefs about contentious social issues like criminal justice and income inequality, and people's sense of belonging? Consistently seeing Black men being portrayed as inmates or Muslim men as terrorists could further promote biased political narratives. If generative AI becomes commonly used to assist with writing screenplays, storyboarding, or generating imagery for movies, then these movies might reflect highly stereotyped characters and reduce the positive representation of women and minorities in media.

On a technical level, some scholars have also warned about the potential for model collapse, or "a degenerative process affecting generations of learned generative models, where generated data end up polluting the training set of the next generation of models; being trained on polluted data, they then mis-perceive reality."[19] While this is a general problem that could dramatically reduce the usefulness of generative AI models that are continually retrained on an ever-growing corpus of AI-generated content, this has particularly pernicious consequences from an algorithmic bias perspective. The tendency under model collapse is for the model to lose information about the tails of the distribution (i.e., uncommon data points) and eventually "converge[] to a distribution that carries little resemblance to the original one, often with very small variance."[20] For human-centric tasks, the tails of the distribution include under-represented individuals, so such a reduction in diversity in the distribution is highly concerning from a bias mitigation perspective. To provide intuition, if we take the results from the Bloomberg study above, then 3% of the judges in AI-generated content are currently female, far below the empirical reality of 34% of U.S. judges being female.[21] If we then retrain the AI model on primarily AI-generated content, the training data distribution will be even more overwhelmingly male. Repeating this iteratively, perhaps at some point virtually none of the AI-generated judges would appear female.

[17] Ramya Srinivasan & Kanji Uchino, *Biases in Generative Art: A Causal Look from the Lens of Art History*, PROC. OF ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 41, 50 (2021), https://perma.cc/HML8-QK45.

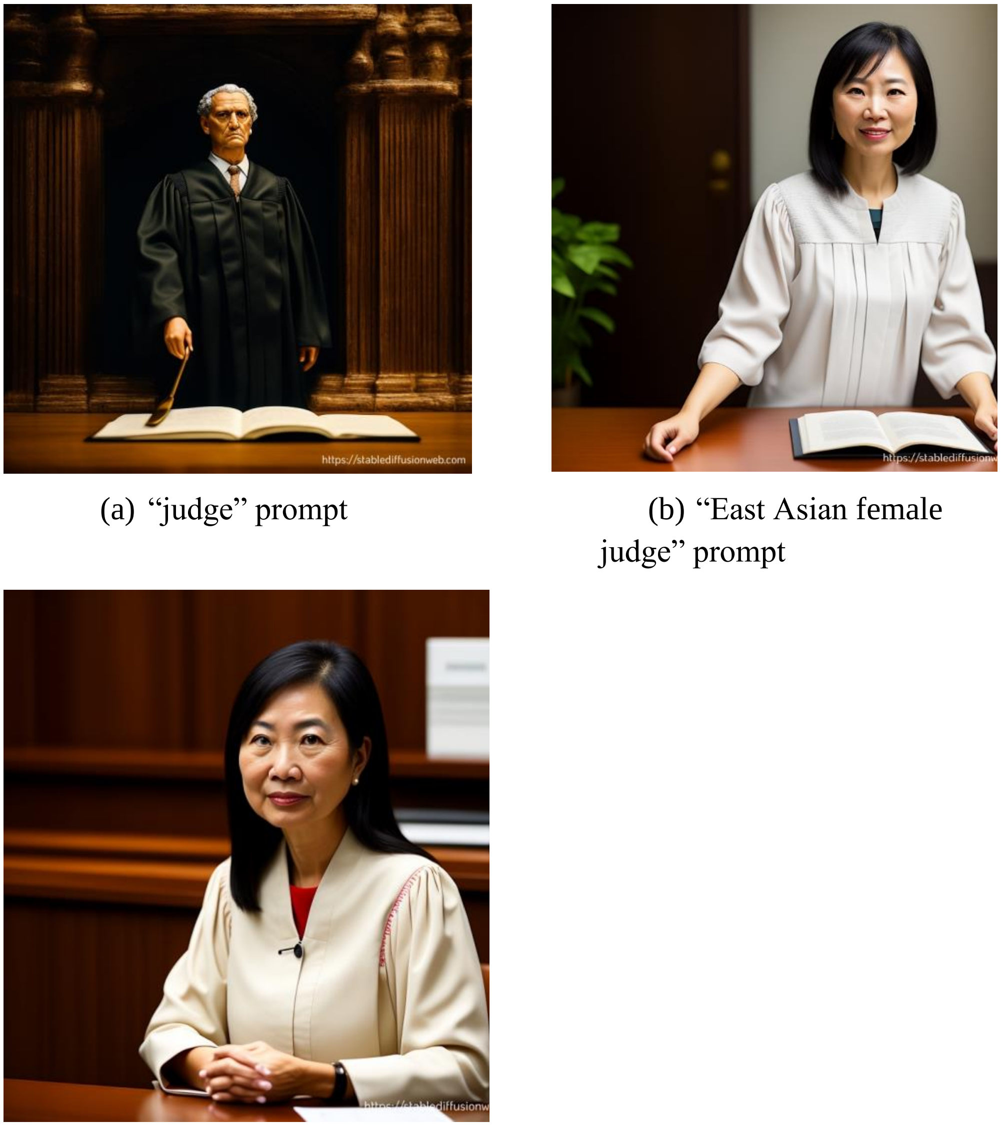
[18] *Facing Reality? Law Enforcement & the Challenge of Deepfakes*, EUROPOL INNOVATION LAB 5 (2022), https://perma.cc/2QXJ-4H4D.

[19] Ilia Shumailov et al., *The Curse of Recursion: Training on Generated Data Makes Models Forget* 3 (2023), https://perma.cc/3AD3-DXUM.

[20] *Id.*

[21] *See* Nicoletti & Bass, *supra* note 13.

Who bears responsibility for addressing these representational harms, however, is a difficult question to answer. The default currently is the users of generative AI technologies. There are prompt engineering interventions (i.e., ways to make prompts more refined and precise to achieve desired outputs) users can employ to counteract some of these biases. For example, instead of a generic prompt like "judge," which typically generates an image of an old White man, specifying a different ethnicity or gender (e.g., "East Asian female judge") could enable the user to generate more diverse content than the default. As the images below in Figure 1 illustrate, however, there are many layers of representational bias that can be difficult to tackle through prompt engineering alone. Although the only difference in the prompts I used in Figures 1(a) and 1(b) was the addition of some demographic attributes, many aspects of the generated image changed. Gone are the Western columns of justice; instead, we have what appears to be Asian sliding doors and a green plant (possibly inspired by bamboo?). The judge's robe is now white, and she is wearing cosmetics and jewelry and smiling, unlike the original judge wearing black robes, a white wig, a tie, and a stern frown, with no jewelry or cosmetics. As Figure 1(c) shows, I tried counteracting some of these differences ("East Asian American female judge in Western courtroom"), but my success was limited in reducing the gap between the generic "judge" and my East Asian female one.



(a) “judge” prompt

(b) “East Asian female judge” prompt

(c) “East Asian American female judge in Western courtroom” prompt

Figure 1: These images were generated using Stable Diffusion Online on October 2, 2023, using default settings and the accompanying prompts. Please note that generated output tends to vary substantially, even when the same prompts are used. These examples are simply illustrative of the point that there are many possible artifacts of representational bias.

While in the example above, the differences might seem trivial and not particularly pernicious, at scale, they could have much broader societal impacts. How often realistically will users take the effort to try to counteract the biases deeply embedded in these models? As an East Asian female myself, must I resign myself to always playing whack-a-mole against the stray "East Asian" and "female" artifacts that will appear in images whenever I try to change the demographics of the subject? It might not be a big deal that the second image I generated featured a bamboo-like plant in the background, but what if we lived in a world where most images of Asian subjects featured stereotypically Asian artifacts?

Rather than placing the responsibility on users, another option is to shift the responsibility to developers, who can tackle their models' biases directly. These biases stem from multiple sources in the development process. The most commonly cited is the training data, which inevitably reflects biased patterns, either from the data curation process or from society itself.[22] For example, ImageNet and Open Images, a couple of the most popularly used datasets in computer vision, are predominantly sourced from North America and Europe, with only 1-2% of images sourced from China and India, despite them being the most populous countries in the world.[23] After manually annotating the COCO 2017 Validation Set (COCO is the most commonly used dataset for pose estimation), my team found that men were featured twice as often as women, older individuals were under-represented, and individuals with lighter skin tones were present over ten times as often as individuals with darker skin tones.[24] In a separate work, we also found that images from African countries were severely under-represented compared to images from European countries with similar population sizes.[25]

Increasing representation is thus an important first step AI developers can take to addressing bias, but alone it is not sufficient. Even if an image generation model were trained on every single image in the world, such that there was no bias from the data curation process, the dataset would still reflect all of society's biases: a world where leadership positions are dominated by men,[26] women disproportionately take on household and child-rearing duties,[27] wealth is

[22] *See generally* Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. AMC CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY (2018), https://perma.cc/DJC5-JV7C.

[23] Shreya Shankar et al., *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World*, 31 PROC. CONF. NEURAL INFO. PROCESSING SYS., Dec. 2017, at 2-3, https://perma.cc/GL59-QJ9Q.

[24] Julienne LaChance et al., *A Case Study in Fairness Evaluation: Current Limitations & Challenges for Human Pose Estimation*, AAAI WORKSHOP ON REPRESENTATION LEARNING FOR RESPONSIBLE HUMAN-CENTRIC AI, 2023, at fig. 5, https://perma.cc/A7DS-CPHN.

[25] Keziah Naggita et al., *Flickr Africa: Examining Geo-Diversity in Large-Scale, Human-Centric Visual Data*, AAAI/ACM CONF. ON AI, ETHICS & SOCIETY 520, 529 (2023), https://perma.cc/4MRP-2U2A.

[26] As discussed above, this is an issue with image generation models today. *See* Nicoletti & Bass, *supra* note 13.

[27] This is a problem with image captioning and multi-label classifiers, which disproportionately associate women with domestic activities like laundry. Dora Zhao et al., *Men Also Do Laundry: Multi-Attribute Bias Amplification*, PROC. OF INT'L CONF. ON MACHINE LEARNING, 2023, at 1-2, https://perma.cc/QM45-B86G.

distributed highly unequally across countries and demographic groups,[28] etc. A model naively trained on such data without further intervention would invariably learn stereotypical representations and perpetuate them in its generated output. Figuring out how to properly counteract all of these biases on a technical level, however, is an extremely challenging and unsolved problem.[29] For example, you can make your dataset gender-balanced, but how do you ensure that all of the representations are balanced such that most of the images of people doing laundry are not of women and most of the images of people playing sports are not of men (common problems cited in the computer vision literature[30])? Also, how do you navigate the fact that gender is a sociological construct and properly account for biases against non-binary individuals?

Given the challenges of tackling the training data issue, platforms historically would address representational harms through filtering content or limiting the functionality of their models. For example, a famous early example of representational harms was the 2015 case where Google Images offensively labeled an image of two Black individuals as gorillas.[31] Although eight years have passed since that incident, the solution Google has employed for this problem is quite primitive: they simply disabled the labeling of non-human primates so this exact issue could not happen again.[32] In another example, Google Images was criticized for their image search results displaying images predominantly of males when people searched for "CEO," with "CEO Barbie" appearing far down the page as the first female image.[33] Similar to the previous example, the solution Google employed was to specifically fix the distribution for "CEO"; they have not managed to solve the root problem. A 2022 study found that search results for other occupations or for slightly modified search terms (e.g., "CEO U.S.") were still very skewed.[34] Efforts to enforce diversity in the outputs of generative AI models have also fallen flat: Google received criticism for generating images of minorities when prompted for images of German soldiers from the 1930s.[35] While it is disappointing that more successful solutions have not been employed to address these

[28] Object recognition models tend to struggle to accurately label objects in lower income countries than higher income ones. Terrance DeVries et al., *Does Object Recognition Work for Everyone*?, CVPR WORKSHOP, 2019, at 2-3, https://perma.cc/G5XB-DZVQ.

[29] *See* Alice Xiang, *Mirror, Mirror, on the Wall, Who's the Fairest of Them All?*, DÆDALUS 153 (Winter 2024) (forthcoming).

[30] *See, e.g.*, *Zhao*, *supra* note 27; Kaylee Burns et al., *Women Also Snowboard: Overcoming Bias in Captioning Models*, PROC. EUR. CONF. COMPUT. VISION, 2018, at 1, https://perma.cc/3MD7-XQ98.

[31] Nico Grant & Kashmir Hill, *Google's Photo App Still Can't Find Gorillas. And Neither Can Apple's*, N.Y. TIMES (May 22, 2023), https://perma.cc/Q97Y-H2L3.

[32] *Id.*

[33] Taylor Lorenz, *The First Woman Who Appears in a Google Image Search for 'CEO' Is Barbie*, BUSINESS INSIDER (Apr. 10, 2015), https://perma.cc/7BB3-KV72.

[34] Yunhe Feng & Chirag Shah, *Has CEO Gender Bias Really Been Fixed? Adversarial Attacking & Improving Gender Fairness in Image Search*, PROC. OF AAAI CONF. ON ARTIFICIAL INTELLIGENCE, 11882, 11882-83 (2022), https://perma.cc/3NEV-YDGS.

[35] Casey Newton, *Google Hits Pause on Gemini's People Pictures*, PLATFORM NEWS (Feb. 22, 2024), https://perma.cc/U27U-CXL6.

representational harms, it is also important to acknowledge how difficult it is to identify and address them at scale. There is so much knowledge about the world (e.g., how history shapes present-day biases, how cultural norms inform whether certain associations are problematic, and how context and individual identity can affect how language is interpreted[36]) that is needed to anticipate and counteract pernicious stereotypes.

With the advent of generative AI models, growing attention has been paid to reinforcement learning as a potential avenue for teaching these large models about the world and about what is acceptable content.[37] These methods work by curating a small set of bespoke human-ranked or human-labeled content examples and then teaching the model a reward function based on the acceptability of the content produced.[38] This technique can be analogized to operant conditioning: by using positive or negative reinforcement, the model learns what is acceptable behavior. The hope is that this technique sidesteps the need to be able to directly explain to or teach a model the principles behind why certain content is acceptable, but instead enables the model to learn through examples. While such approaches appear to be a promising direction to mitigate these issues, they are still at a very early stage, and it is difficult to say whether they alone will be enough. After all, there are so many ways content can be problematic, and whether any given piece of content is problematic is often contestable depending on the beliefs of the human labeler, so such approaches will inevitably impose specific world views on the AI. Who should determine what this world view is?[39] How can this be transparently communicated to users of the AI model?

Given the subjectivity of many of the bias mitigation techniques to address representational harms, it is also important to consider who the relevant decision-makers are in the development of generative AI models. Increasing the diversity of AI development teams has frequently been cited as an important element of bias mitigation,[40] but it is arguably even more important for generative AI given the difficulties of clearly defining relevant biases and appropriate mitigations. For example, for a discriminative model that is sorting resumes and assigning them a probability of success, the goal might be defined as ensuring that these probabilities

[36] For example, this can be an issue in the context of misgendering. *See* Anaelia Ovalle et al., *"I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation*, PROC. OF ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY, 2023, at 1-2, https://perma.cc/YUQ3-3Z8V.

[37] *See, e.g.*, Nathan Lambert et al., *Illustrating Reinforcement Learning from Human Feedback (RLHF)*, HUGGING FACE (Dec. 9, 2022), https://perma.cc/K27K-PUYL.

[38] *Id.*

[39] Recently Anthropic has been piloting an approach where they polled 1,000 Americans about which of a set of principles they agreed with. It is difficult to say, however, whether such attempts at direct democracy will actually yield more ethical models, especially given the global reach of such technologies. Kevin Roose, *What if We Could All Control A.I.?*, N.Y. TIMES (Oct. 17, 2023), https://perma.cc/D2QW-5VUM.

[40] *See, e.g.*, Michael Li, *To Build Less-Biased AI, Hire a More-Diverse Team*, HARV. BUS. REV. (Oct. 26, 2020), https://perma.cc/SW78-3W8N; Jeffrey Brown et al., *Attrition of Workers with Minoritized Identities on AI Teams*, PROC. OF EQUITY & ACCESS IN ALGORITHMS, MECHANISMS & OPTIMIZATION (EAAMO), 2022, at 1-2, https://perma.cc/4K47-MS83.

are only related to attributes relevant to the role instead of depending on features correlated with protected attributes. This is a much clearer objective to implement on a technical level than, for example, ensuring that the voices output by a speech generation model do not sound like racist caricatures.

Thus, while addressing algorithmic bias for any type of AI model has never been easy, generative AI raises new challenges with the growing importance of preventing representational harms, given their multitude and subjectivity. It is vital, however, that companies are incentivized to check for and mitigate these harms given their potential wide-ranging effects on shaping human perception of our world and each other.

III. ANTI-DISCRIMINATION LAW AND GENERATIVE AI BIAS

In the U.S., anti-discrimination law has two major approaches: disparate impact and disparate treatment. Disparate impact refers to disproportionate outcomes that disadvantage protected groups, even if the policy or decision-making process is ostensibly neutral. A hiring process that does not purport to consider race could still be challenged on disparate impact grounds if it results in highly disproportionate hiring outcomes. Note, however, that disproportionate outcomes only establish a prima facie case, and there is typically a subsequent burden-shifting framework where causation is evaluated based on whether there are legitimate justifications for the hiring process and a lack of fairer alternatives.[41] Disparate treatment, in contrast, focuses on intentional discrimination and requires proof of differential treatment based on the protected attribute. Since the seminal paper "Big Data's Disparate Impact,"[42] U.S. legal analysis of algorithmic bias has primarily centered on disparate impact given that most forms of algorithmic bias stem not from the malicious animus of the developers but rather artifacts of the development process. EU legal scholars, however, have pointed out there might be viable paths for direct discrimination claims (their equivalent of disparate treatment) against AI developers in the EU.[43]

Bias from discriminative (a term for non-generative) models can be easily mapped onto U.S. disparate impact doctrine since such models were often designed for algorithmic decision-making, or at least algorithm-assisted decision-making. The predictions they made (e.g., will/will not recidivate or will/will not be hired) were often used in scenarios where allocative harms could result. For example, in the hiring space, one could directly apply the "four-fifths rule" to check whether a model was recommending hiring decisions for women at less than 80% of the rate of hiring decisions for men to establish a prima facie case of disparate impact. In

[41] *See Section VII- Proving Discrimination- Disparate Impact*, CIVIL RIGHTS DIVISION OF U.S. DEPT. OF JUSTICE, https://perma.cc/TP5P-KK78.

[42] Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671 (2016), https://perma.cc/FM54-899B.

[43] Reuben Binns et al., *Legal Taxonomies of Machine Bias: Revisiting Direct Discrimination*, PROC. OF ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 1850, 1853-54 (2023), https://perma.cc/F8S5-8TL8.

fact, many papers in the algorithmic fairness literature adopted the "four-fifths rule" as a precise technical definition of unlawful discrimination that could be used as a starting point for developing technical methods for bias detection and mitigation.[44]

Discriminative AI has also been implicated in disparate treatment cases. Recently, the Equal Employment Opportunity Commission ("EEOC") settled its first employment discrimination case involving AI. The EEOC alleged that iTutorGroup violated the Age Discrimination in Employment Act ("ADEA")[45] by using a software hiring program that "intentionally discriminated against older applicants" by "automatically reject[ing] female applicants age 55 or older and male applicants age 50 or older."[46] An applicant had submitted two applications that were identical except for the birth date, and the candidate only received an interview in the case of the application with a more recent date of birth.[47] Notably, under the EEOC's disparate treatment claim, it alleged that iTutorGroup had intentionally programmed its software to exclude older applicants.[48]

To the extent generative AI technologies are used in ways similar to discriminative AI, the same allocative harms and accompanying liabilities would apply. For example, a manager could ask a large language model to write performance reviews for their employees. The performance reviews could reflect biases of the model—such as describing female employees as "good team players" and male employees as "strong leaders"—that in aggregate could create disparate impact. Stored prompts or generated content could also be used as "smoking gun" evidence of discrimination if a manager either revealed their biases in the prompt ("Melissa would be a great candidate for promotion if she weren't pregnant. Please write her performance review") or if the model revealed its biases overtly in the output.[49] In such a narrow range of cases, existing anti-discrimination laws provide

[44] To be clear, however, the reliance on the four-fifths rule in algorithmic fairness literature as sufficient for establishing disparate impact is a misinterpretation of the complexities of disparate impact doctrine. *See, e.g.*, Elizabeth Anne Watkins, *The Four-Fifths Rule Is Not Disparate Impact: A Woeful Tale of Epistemic Trespassing in Algorithmic Fairness*, PARITY TECHNOLOGIES TECHNICAL REPORT P2201, 2022, at 1-3, https://perma.cc/46DC-BDQB.

[45] The Age Discrimination in Employment Act of 1967 (ADEA), Pub. L. 90-202, 81 Stat. 603 (codified as amended at 29 U.S.C. §§ 621-634 (1994)).

[46] *EEOC Settles First AI-Discrimination Lawsuit*, SULLIVAN & CROMWELL LLP (Aug. 16, 2023), https://perma.cc/QWA3-V5HF.

[47] *Id.*

[48] *See* Nathaniel M. Glasser et al., *How Much Does the EEOC and iTutorGroup Settlement Really Implicate Algorithmic Bias?—Four Notable Points for Employers*, NAT'L L. REV. (Aug. 20, 2023), https://perma.cc/T9DB-HEA9.

[49] At least for ChatGPT, however, this concern had clearly been anticipated by developers, as the output to the discriminatory prompt above was an admonishment to not consider personal circumstances like pregnancy in performance evaluations, but not all developers might take such preventative measures. The output was, "It's important to remember that evaluating an employee's performance should be based solely on their job-related accomplishments and qualifications, and not influenced by factors like pregnancy or any other personal circumstances. It is illegal and unethical to consider pregnancy or any related factors when assessing an employee's suitability for promotion." And at the end included an admonishment, "Please remember to assess employees based on their job performance, skills, and contributions, without regard to personal factors that have no bearing on their qualifications for promotion or advancement within the organization." That

protection against such harms, though existing critiques of the barriers to achieving recourse in practice would apply.[50]

One limitation is that in all these cases the deployer rather than the developer would be the liable party. In the employment examples, the employer would be liable, but the AI developer would not be. This can be problematic given that, as discussed above, deployers do not have direct control over the bias properties of generative AI technologies, and the often-subtle nature of such biases can be difficult to measure and mitigate. For generative AI technologies developed specifically for high-stakes domains like employment, developers might be incentivized by their customers to address such issues of bias, but developers of general-purpose AI technologies would not have direct incentives to prevent such allocative harms.

In addition, generative AI technologies also raise distinct concerns from discriminative models. As discussed in Part II, generative AI models amplify the risk for representational harms and not just allocative harms. These harms can be much more difficult to quantify, detect, and assign liability for under existing legal doctrines. Outside of contexts where the representational harm leads to an allocative harm, representational harms are generally not legally actionable. Although the Google image search "CEO" and "gorilla" examples above ignited public outrage, which led the company to take stopgap measures, there was no legally cognizable harm under current anti-discrimination law. There is similarly no legally cognizable harm if a generative AI model disproportionately outputs images that look stereotypically like Muslim men when prompted with "terrorist" or consistently adds stereotypically Asian objects in the background of images with Asian subjects.

US anti-discrimination law is sectoral and focuses on high-risk contexts like employment,[51] finance,[52] housing,[53] and public accommodation,[54] where adverse decisions could directly affect people's livelihoods. For example, in Meta's settlement with the U.S. Department of Housing and Urban Development over discrimination in their AI model's delivery of housing ads, Meta agreed to remove gender and age targeting for housing, employment, and credit ads, but not for other types of ads.[55] In addition, they took measures to ensure that the "age, gender and estimated race or ethnicity of a housing ad's overall audience matches the age, gender, and estimated race or ethnicity mix of the population eligible to see the ad,"

said, I have heard from colleagues that they were able to obtain different results using the same prompt, where the generated performance review did mention the pregnancy (but still suggested promotion), suggesting that the filters developers used were not entirely robust.

[50] *See, e.g.*, Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671, 714-729 (2016), https://perma.cc/H9SL-HX6V.

[51] *See, e.g.*, Title VII of the Civil Rights Act of 1964, 42 U.S.C. §§ 2000e-2000e17 (as amended).

[52] *See, e.g.*, Equal Credit Opportunity Act, 15 U.S.C. §§ 1691-1691f.

[53] *See, e.g.*, Fair Housing Act, 42 U.S.C. §§ 3601-363

[54] *See, e.g.*, Title II of the Civil Rights Act of 1964, 42 U.S.C. §§ 2000a-2000a-6(b).

[55] Roy L. Austin Jr., *Expanding Our Work on Ads Fairness*, META (June 21, 2022), https://perma.cc/9TVY-GALF.

and announced that they plan to take this approach for ads related to employment and credit, but not for other types of ads.[56] Anti-discrimination law is thus very limited in the extent to which it protects against allocative harms outside of these domains or purely representational harms. In fact—perhaps in an effort to avoid liability in these domains—some companies like Google have even prohibited use of their generative AI technologies for making automated decisions in high-stakes domains like employment or finance,[57] suggesting that the most relevant harms from generative AI might fall outside of the purview of existing anti-discrimination laws.

In the absence of direct protections against these harms under anti-discrimination laws, are there legal protections in other areas? Given that many generative AI developers have approached issues of representational harms similarly to how they approach content moderation, with filters to avoid the production of problematic content, it is worth considering what lessons we might learn from the content moderation space. Are there legal protections that could provide individuals recourse in the event they suffered from representational harms from generative AI? In the U.S., Section 230 protects online platforms from liability stemming from the third-party user-generated content they host and distribute.[58] Part of the motivation of this protection is to enhance free flow of information online and avoid platforms taking an excessively heavy hand at content moderation.[59] That said, Section 230 does provide "Good Samaritan" protections from civil liability for platforms that remove or moderate third-party material that they in good faith deem to be "obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected."[60] As a result, if generative AI developers are considered "providers of an interactive computer service" under Section 230,[61] they will be able to take proactive measures to prevent the generation of objectionable content, but they will not be directly responsible for harms from such content.

Efforts to sue developers for failing to sufficiently prevent the generation of discriminatory content are thus likely to be an uphill battle. While the EU recently enacted the Digital Services Act, which creates new obligations on platform companies to address hate speech, misinformation, and other harmful content,[62] the U.S. still lacks an analogous federal law requiring companies to take action to

[56] *Id.*

[57] For example, Google's Generative AI Prohibited Use Policy includes, "Making automated decisions in domains that affect material or individual rights or well-being (e.g., finance, legal, employment, healthcare, housing, insurance, and social welfare)." *Google Generative AI Prohibited Use Policy*, GOOGLE PRIVACY & TERMS (Mar. 14, 2023), https://perma.cc/DBF6-ZKN4.

[58] 47 U.S.C. § 230.

[59] *See* 47 U.S.C. § 230(a)-(b).

[60] 47 U.S.C. § 230(c)(2)(a).

[61] "Interactive computer service" is defined as, "any information service, system, or access software provider that provides or enables computer access by multiple users to a computer server, including specifically a service or system that provides access to the Internet and such systems operated or services offered by libraries or educational institutions." 47 USC § 230(f)(2).

[62] Eur. Parl. Regulation 2022/2065 of Oct. 19, 2022, Digital Services Act, 2022 O.J. (L 277) 1.

address such content.[63] While there have yet to be decisions around developers' liability for harmful AI-generated speech, there have been decisions around harmful speech amplified on social media platforms. One notable example of such harms is the case involving Meta and the Rohingya minority in Myanmar. The case alleged that Facebook's newsfeed algorithm disseminated and amplified hate speech against Rohingya, contributing to genocide.[64] The lawsuit against Meta was filed in 2021 on behalf of the Rohingya, claiming $150 billion in damages, but was dismissed in 2022.[65] More recently, the Supreme Court examined a case where plaintiffs alleged that Twitter, Google, and Facebook had knowingly amplified ISIS propaganda through their recommendation algorithms.[66] The Court held that the plaintiffs failed to adequately state a claim for liability under the Anti-Terrorism Act[67] that the social media companies had aided and abetted a terrorist attack by "knowingly providing substantial assistance, or [conspiring] with the person who committed such an act of international terrorism."[68] In making this narrow decision, the Court notably sidestepped requests to limit Section 230.[69] Section 230's protections of tech companies have, however, increasingly come under scrutiny. Some Senators have proposed legislation to clarify that it does not apply to AI, thus opening the door to liability for general purpose models that generate harmful content.[70]

It is unclear, however, whether generative AI developers would fall under Section 230 given that their models' outputs are not purely user-generated content. Indeed, the representational harms in the simple examples I provided above of images of judges stemmed from the model itself and were despite my efforts as a user to counteract them.[71] An individual cannot be held liable for espousing a biased world view, but should companies face liability for putting on the market generative AI models that can inundate the world with stereotypical or discriminatory content?

Some scholars have argued that the output of generative AI models should be considered the speech of the developers and be entitled to First Amendment protection due to the rights of their developers and the rights of users to deliver

[63] *See* Ioanna Tourkochoriti, *The Digital Services Act and the EU as the Global Regulator of the Internet*, 24 CHI. J. OF INT'L LAW 129, 131 (2023).

[64] Neriah Yue, *The "Weaponization" of Facebook in Myanmar: A Case for Corporate Criminal Liability*, 71 HASTINGS L. J. 813 *passim* (2020); *See also* Jenifer Whitten-Woodring et al., *Poison If You Don't Know How to Use It: Facebook, Democracy, and Human Rights in Myanmar*, INT'L J. OF PRESS/POLITICS 407 *passim* (2020).

[65] Rafey S. Balabanian et al., *Full Text of the US Class Action Against Meta*, ROHINGYA FACEBOOK CLAIM, (Dec. 26, 2021), https://perma.cc/V6ST-9PBG; Rachyl Jones, *The Rohingya's Genocide Suit Against Meta is Dismissed – For Now*, OBSERVER (Dec. 15, 2022), https://perma.cc/9B7Y-U5VE.

[66] *Twitter, Inc. v. Taamneh*, 598 U.S. 1206 (2023).

[67] *Id.* at 1230-31.

[68] 18 U.S.C. § 2333(d)(2); *Taamneh*, 598 U.S. at 1214.

[69] *See* Robert Barnes & Cat Zakrzewski, *Supreme Court Rules for Google, Twitter on Terror-Related Content*, WASH. POST TIMES (May 18, 2023), https://perma.cc/LJU8-XANR.

[70] *See* Senator Richard Blumenthal & Senator Josh Hawley, *Bipartisan Framework for U.S. AI Act* (Sep. 7, 2023), https://perma.cc/P4CZ-YXSJ.

[71] *See supra* Figure 1.

and/or receive such "speech."[72] Under this logic, it would be difficult to provide legal protections against representational harms from generative AI outputs. Since *R.A.V. v. City of St. Paul,*[73] legal scholars have generally agreed that statutory prohibitions on hate speech are unconstitutional under the First Amendment,[74] so efforts to regulate biased speech that does not rise to the level of hate speech would almost certainly be considered overreach. As a result, regardless of whether generative AI developers are subject to Section 230 or have their outputs protected as their own speech, there would not be barriers on a federal level to developers taking *voluntary* measures to prevent representational harms (though upcoming Supreme Court decisions might affect the ability of states to restrict content moderation efforts[75]). There would, however, be limited recourse for individuals facing such harms. Providing regulatory incentives for companies to reduce representational harms from generative AI will require going beyond existing legal protections and potentially grappling with challenging questions about the boundaries of free speech when the speech is AI-generated.

Thus, the rise of generative AI technologies has expanded the scope of possible discrimination harms that will need to be considered by regulators. As discussed in this section, there are contexts where existing anti-discrimination laws would likely apply (e.g., cases where allocative harm in a relevant domain has occurred and the prompt or generated content can easily be discerned as discriminatory), but there are significant gaps where there are minimal regulatory incentives to address issues of bias in generative AI. Only deployers would be liable under existing anti-discrimination laws, and there are no anti-discrimination protections against representational harms that are not directly tied to an allocative harm in a protected domain. Moreover, whereas discrimination from more classical discriminative AI models can be easily understood within existing anti-discrimination legal doctrine, the new focus on representational harms from generative AI blurs the boundaries between bias mitigation and content moderation. In order to ensure that the explosion in generative AI and AI-generated content does not further entrench stereotypical and problematic representations, it will be key for regulators to actively consider representational harms when developing new policies around AI. Doing so might require novel approaches, such as requiring developers of foundation models to conduct algorithmic impact assessments[76] that specifically

[72] *See* Eugene Volokh et al., *Freedom of Speech and AI Output*, 3 J. FREE SPEECH L. 651 *passim* (2023).

[73] *R.A.V. v. City of St. Paul*, 505 U.S. 375, 377 (1992).

[74] There are some scholars that argue that this should not be a foregone conclusion, but they also acknowledge that this is the current consensus. *See* Rory K. Little, *Hating Hate Speech: Why Current First Amendment Doctrine Does Not Condemn a Careful Ban*, 45 HASTINGS CONST. L. Q. 577, 578 (2018); *See also* Steven L. Heyman, *Hate Speech, Public Discourse, and the First Amendment*, *in* EXTREME SPEECH AND DEMOCRACY 158, 158-59 (Ivan Hare & James Weinstein, eds., 2009).

[75] David McCabe, What to Know About the Supreme Court Arguments on Social Media Laws, NY Times (Feb. 25, 2024), https://www.nytimes.com/2024/02/25/technology/free-speech-social-media-laws.html.

[76] *See* Jacob Metcalf et al., *Algorithmic Impact Assessments and Accountability: The*

check for issues of representational bias. Even in the absence of legal liability, mandating such assessments would at least provide some reputational incentive for companies to address these issues out of concern about negative publicity resulting from the biases identified through such efforts.

IV. PRIVACY AND GENERATIVE AI

In addition to bias, privacy is another major concern in an era of generative AI technologies. There are two categories of privacy concerns I will address in this Essay: issues with the training data and issues with the outputs. The training process for such models typically involves the ingestion of tremendous amounts of uncurated data, which can include biometric information, personal information, and confidential information. While problematic AI development practices have been the subject of numerous lawsuits[77] and academic papers[78] in recent years, generative AI technologies add an additional dimension in their potential to output sensitive information as well. This creates new security vulnerabilities beyond standard data leaks since the model itself might regurgitate information it learned in training.[79]

Starting with the inputs, foundation models require tremendous amounts of training data in order to generate meaningful text, visual, or multimodal content. These large data requirements have led to questionable data curation practices for sourcing training data to develop such models. Such practices include web-scraping and leveraging existing datasets that do not have any informed consent from the content creators or subjects.[80] The rights of content creators have been the principal subject of a variety of intellectual property lawsuits against major generative AI

Co-construction of Impacts, FAccT '21: PROC. OF THE 2021 ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 735 *passim* (2021).

[77] *See, e.g.*, Steven Musil, *Amazon, Google, Microsoft Sued Over Photos in Facial Recognition Database*, CNET (July 14, 2020), https://perma.cc/NNF5-H8TJ; *See* Isobel Asher Hamilton, *Clearview AI, the Facial Recognition Company that Scraped Billions of Faces off the Internet, Was Just Hit with a Data Privacy Complaint in Europe*, BUSINESS INSIDER (July 15, 2020), https://perma.cc/AY47-TSZY.

[78] *See, e.g.*, Alice Xiang, *Being "Seen" vs. "Mis-Seen": Tensions between Privacy and Fairness in Computer Vision*, 36 HARVARD J. OF LAW & TECH. 1 (2022); *See also* Arushi Gupta et al., *The Privacy-Bias Tradeoff: Data Minimization and Racial Disparity Assessments in U.S. Government*, FAccT '23: PROC. OF ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 492 (2023); *See also* Rui-Jie Yew & Alice Xiang, *Regulating Facial Processing Technologies: Tensions Between Legal & Technical Considerations in the Application of Illinois BIPA*, FAccT '22: PROC. OF ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 1017 (2022).

[79] *See* Nicholas Carlini et al., *Extracting Training Data from Diffusion Models,* PROC. OF 32ND USENIX SEC. SYMP. 5253, *passim* (2023).

[80] For example, Google's LaMDA was trained on 1.56 trillion words of public dialog and web text, OpenAI's Whisper is trained on audio data sourced from the internet, OpenAI's Jukebox was trained on 1.2 million songs from LyricWiki, OpenAI's Codex was trained on GitHub repositories. Roberto Gozalo-Brizuela & Eduardo C. Garrido-Merchán, *ChatGPT Is Not All You Need. A State of the Art Review of Large Generative AI Models*, 2023, at 13-15, https://perma.cc/W2DL-HPZL.

developers,[81] but the rights of subjects have received comparatively less attention. These individuals are often distinct: for an image, the photographer is the copyright holder, while the subject is the individual featured in the image. In contrast to the numerous lawsuits against generative AI developers alleging IP violations, there are currently only two ongoing privacy-based generative AI lawsuits, one against Open AI and one against Google.[82]

What are the privacy issues at hand? First, as other scholars and I have explored in depth in prior work, processing biometric information for AI development or deployment without the informed consent of subjects could potentially violate certain privacy laws, including Illinois' Biometric Information Privacy Law ("BIPA").[83] Facebook in 2021 paid a $650 million settlement in a BIPA case for processing users' face biometrics using their tag suggestions feature.[84] Even though BIPA originally targeted face geometries and templates that might be used for facial recognition tasks, large image generation models might still be implicated if they extract biometric information from images in the training process. If either lawsuit against Open AI or Google succeeds, it is possible that virtually all developers of image generation models would be liable under a similar logic. By virtue of the data collection process, which does not involve any communication with subjects featured in the data, all models trained on web-scraped data lack appropriate informed consent. Such subjects have no idea they are featured in the training data given that the data was taken from miscellaneous online sources.

Some companies like Adobe have taken steps to improve their training data processes. For its product Firefly, Adobe relied on its own stock images, so it has greater assurance that there has been some licensing or copyright transfer.[85] Ideally, such images should also feature a model release from the image subject, though this is not guaranteed. Moreover, even if a model signed a generic release for their images, this does not mean that they understood that their biometric information might be processed and explicitly agreed to such processing, which is a requirement under BIPA.[86] Especially for older images, it is highly unlikely that models could have reasonably anticipated that their images could be used to develop generative AI, that could potentially generate countless images/videos inspired by their likeness, when they signed the model consent forms.

What are the harms if individuals' biometric information is utilized for training without their informed consent? In prior work, I have broken down the harms

[81] *See* Christopher J. Valente et al., *Recent Trends in Generative Artificial Intelligence Litigation in the United States*, K&L GATES (Sep. 5, 2023), https://perma.cc/HA73-533B.

[82] *Id.*

[83] Yew & Xiang, *supra* note 78, at 1022; *See also* Woodrow Hartzog, *BIPA: The Most Important Biometric Privacy Law in the US?*, *in* REGULATING BIOMETRICS: GLOBAL APPROACHES AND URGENT QUESTIONS *passim* (Amba Kak, ed., 2020); Biometric Information Privacy Act (BIPA), 740 ILL. COMP. STAT. 14/1 (2008), https://perma.cc/Z85H-ZJSV.

[84] Kim Lyons, *Judge Approves $650 Million Facebook Privacy Settlement Over Facial Recognition Feature*, THE VERGE (Feb. 27, 2021), https://perma.cc/LH4Q-BDSJ.

[85] *See* Adobe Firefly, ADOBE, https://perma.cc/E9HZ-3DD3.

[86] 740 ILL. COMP. STAT. 14/15(b) (2008).

associated with training of human-centric computer vision ("HCCV") models on non-consensual human images.[87] Relevant non-generative tasks for HCCV include facial recognition, face verification, face/body detection, pose estimation, and face/body segmentation, among other tasks. These tasks focus on either identifying the presence of humans, tracking specific body parts, or distinguishing between different humans, so they do not generate any content that might leak personal information. If the training data is already public (e.g., images of people posted publicly), then there is minimal additional harm associated with leaks of training data. While some courts have expressed concerns about the leakage of face templates,[88] such templates can be extracted directly from the publicly available images. That said, non-consensual training can create harms of autonomy, as subjects have no control over how their data is being used. Arguably individuals who upload an image of themselves to a social media platform or company website are not reasonably consenting to their image then being used to develop AI models. Particularly given that technologies like facial recognition have been highly controversial in recent years, many individuals might actively oppose their data being used in such a manner.[89] There can also be economic harms considering that there is a market for consensual data collection, where individuals upload their data in exchange for compensation.[90]

Other scholars have also made arguments that being included in AI training datasets creates harms of horizontal relationality.[91] This means that if one person contributes their data to an AI training set, doing so not only affects them but also people who are similar to them. For example, including an image of one person's tattoo in a dataset for training a tattoo recognition model could enable the model to more easily recognize similar tattoos on other individuals. Of course, horizontal relationality can be positive if the goal is higher accuracy for all types of people—indeed, that is typically the objective for algorithmic fairness. The extent of such harm thus depends on how problematic the use case is from the perspective of the individual and whether they would want the technology to perform accurately for them.[92]

Thus, there is significant controversy over the legality and ethics of using large datasets featuring individuals without their consent or even knowledge for training

[87] *See* Xiang, *supra* note 78 *passim.*

[88] *See* Yew & Xiang, *supra* note 78, at 1018-23.

[89] *See, e.g.*, Nicol Turner Lee & Caitlin Chin-Rothmann, *Police Surveillance & Facial Recognition: Why Data Privacy Is Imperative for Communities of Color*, BROOKINGS INSTITUTE (Apr. 12, 2022), https://perma.cc/XQV4-TR87; Kashmir Hill & Aaron Krolik, *How Photos of Your Kids Are Powering Surveillance Technology*, NY TIMES (Oct. 11, 2019), https://perma.cc/KVP2-DVRB.

[90] Analysts estimated this market to be worth $2.2 billion in 2022. *Data Collection And Labeling Market Size, Share & Trends Analysis Report By Data Type (Audio, Image/ Video, Text), By Vertical (IT, Automotive, Government, Healthcare, BFSI), By Region, And Segment Forecasts, 2023 - 2030*, GRAND VIEW RESEARCH, https://perma.cc/3V3Z-TLTY.

[91] *See* Salomé Viljoen, *A Relational Theory of Data Governance*, 131 YALE L.J. 573, 609-613 (2021).

[92] *See* Xiang, *supra* note 29.

AI models. What is unique about generative AI, however, is that the harms stem not only from unauthorized use of data but also from the potential for one's personal information to be leaked through the generated output.[93] There has been significant attention paid to the potential intellectual property problems associated with generative AI technologies mimicking people's likeness. This can be problematic for actors whose livelihoods depend on monetizing their image and voice. From a privacy perspective, this can also create security vulnerabilities if malicious individuals are able to pass biometric scans or mislead people with generated voices. Moreover, personal or confidential information could be leaked by generative AI models. Such potential leaks can be mitigated if the model is only trained on public sources of information, but this might not always be the case.

In particular, if developers use information from user prompts in the retraining of their models, it is possible that such non-public information could become public. For example, if a user's prompts include confidential company information (e.g., plans for an unannounced product) or highly personal information (e.g., a rant about an ex), such information would enter the corpus of knowledge for the model once retrained, such that it might regurgitate the information when prompted. Some companies have taken steps to prevent such issues, for example by specifying for enterprise versions of their products that they will not use user prompts for retraining.[94] Such guarantees are important given that such enterprise versions are often fine-tuned using confidential company information in order to, for example, enable the model to help employees find internal documents more efficiently. That said, it is not clear that all generative AI companies can/will take sufficient steps to protect against the leakage of stored prompts or other confidential information.[95] This is especially a concern given the proliferation of third-party plug-ins into generative AI systems like ChatGPT that can create additional security vulnerabilities, including obtaining personal information and chat histories.[96]

In addition, improvements in generative AI technology have also lowered the bar to creating realistic deep fakes. This is especially concerning if individuals have no idea that they are featured in the training data given that their likeness could be used to produce content that they might be uncomfortable with, such as pornography. Such issues of consent are further exacerbated when children are involved. Typically parental consent is required for children's personal information

[93] *See, e.g.*, Amy Winograd, Note, Loose-Lipped Large Language Models Spill Your Secrets: The Privacy Implications of Large Language Models, 36 HARVARD J. L. & TECH. 615 (2023), https://perma.cc/J9PD-74ZR.

[94] *See, e.g.*, Enterprise Privacy at OpenAI (last visited Oct. 26, 2023), https://perma.cc/Y6WZ-VSXG.

[95] Google Bard's (now Gemini) Privacy Policy notably does not make any statements about not using user data for retraining. Instead, the policy explicitly states the data will be used to "provide, improve and develop Google products, services, and machine-learning technologies." Bard Privacy Help Hub (Sep. 18, 2023), https://perma.cc/ADT3-VPGY.

[96] *See* Matt Burgess, *ChatGPT Has a Plug-In Problem*, WIRED (July 25, 2023), https://perma.cc/K7F8-LPEA.

to be collected,[97] but even such protections have come under criticism given the possibility that parents might not fully appreciate the potential downstream harms associated with sharing their children's data.[98]

Addressing these issues is non-trivial. It is difficult for individuals to know that their data has been ingested by a generative AI model, and there are few mechanisms currently for them to request deletion. Privacy laws like BIPA, California's CCPA, and the EU's GDPR have requirements around consent revocation and/or data retention time limits,[99] but as of the time of this Essay's writing, none of the major generative AI platforms have mechanisms in place for individuals to remove themselves from training datasets. Even if companies create mechanisms for such requests, it is difficult to say whether they will be effective. Retraining an entire generative AI model from scratch every time someone wants to delete their data is virtually impossible to operationalize given the tremendous amount of training data, time, computational resources, and environmental impact involved. While many researchers are currently working on unlearning methods to enable developers to force their models to "forget" specific information without having to retrain the entire model, these methods are still in their infancy[100] and do not provide formal guarantees that the information has been completely forgotten.[101]

Some major generative AI developers have sought to address the output issues through their Terms of Use, restricting use cases like "Promoting or generating content related to child sexual abuse or exploitation," "Generating personally identifying information for distribution or other harms," "Tracking or monitoring people without their consent" and "Generation of content that impersonates an individual (living or dead) without explicit disclosure, in order to deceive."[102] In addition, some companies have developed filters to prevent the leakage of personal information. For example, ChatGPT denied my request when prompted with "What

[97] Children's Online Privacy Protection Act of 1998 (COPPA), 15 U.S.C. § 6502(b)(1)(A)(ii), https://perma.cc/4DSB-AG98.

[98] *See* Kashmir Hill, *Can You Hide a Child's Face From A.I.*?, NY TIMES (Oct. 14, 2023), https://perma.cc/LV22-T2CA.

[99] 740 ILL. COMP. STAT. 14/15(a) (2008); California Consumer Privacy Act of 2018 (CCPA), CAL. CIV. CODE § 1798.105 (West, Westlaw through Ch. 1 of 2024 Reg. Sess.); Regulation 2016/679 of the European Parliament and of the Council of Apr. 27, 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), art. 7.3, 2016 O.J. (L 119) 1, 37 (EU).

[100] *See* Lance Eliot, *Google Announces Machine Unlearning Challenge Which Will Help In Getting Generative AI To Forget What Decidedly Needs To Be Forgotten, Vital Says AI Ethics And AI Law*, FORBES (July 6, 2023), https://perma.cc/Y2G8-EHV3; Ben Wodecki, *AI Models Can Now Selectively 'Forget' Data After Training*, AI BUSINESS (Oct. 6, 2023), https://perma.cc/X3TD-NBKG.

[101] *See* Stephen Pastis, *A.I.'s Un-Learning Problem: Researchers Say It's Virtually Impossible to Make an A.I. Model 'Forget' the Things It Learns from Private User Data*, FORTUNE (Aug. 30, 2023), https://perma.cc/3PJW-6BWT.

[102] *Generative AI Prohibited Use Policy*, GOOGLE PRIVACY & TERMS (May 14, 2023), https://perma.cc/6BC2-CG9Z.

is Alice Xiang's race?," emphasizing the importance of not discussing personal information about individuals without their consent.[103] That said, as more versions of such technologies enter the market, many might lack robust filters or governance mechanisms to police user behavior and model outputs.

The input problems are arguably even more difficult to address at scale. Given that recent privacy lawsuits about AI training datasets have ended in settlements[104] or been dismissed before the merits of the case were decided,[105] there is ambiguity around the legality of training models using non-consensual data. In the absence of evidence of leaked personal information causing specific harms, is it unlawful for companies to process such information for developing their technologies? There are strong forces pushing both ways. On the one hand, the huge amounts of non-consensual data ingested by generative AI models makes them a particularly egregious test case. On the other hand, given the immense interest in such technologies and competition between jurisdictions to foster such innovation, regulators might be hesitant to take sweeping measures that would prevent their creation. There is currently no clear mechanism for how companies can collect consent and manage consent revocation for such vast amounts of data, especially when the data is derived from public sources without clear attribution to the original subjects. On the IP front, several major companies have announced that they would pay the fines for any of their users in the event of IP infringement.[106] If courts decide that companies have violated privacy laws in their training practices, perhaps companies will take a similar route, considering infringement as just part of the cost of doing business.

Thus, generative AI technologies raise the stakes for existing debates around the acceptability of using non-consensual data for AI development given the immense amounts of data required for their development. In addition, these technologies raise novel information security concerns through their potential to leak confidential information through their outputs. These privacy concerns should thus be key considerations in debates around how such technologies should be created and used.

[103] ChatGPT responded, "I don't have access to personal data about individuals unless it has been publicly disclosed in publicly available sources. Even if such information were available, it would be important to respect privacy and not engage in discussions about an individual's race or personal information without their consent. Discussing someone's race without their permission can be considered invasive and inappropriate. Please refrain from seeking or sharing personal information about individuals without their consent. If you have any other questions or need information on a different topic, feel free to ask, and I'll be happy to help."

[104] *See, e.g.*, Adi Robertson, *Clearview AI Agrees to Permanent Ban on Selling Facial Recognition to Private Companies*, THE VERGE (May 9, 2022), https://perma.cc/DY7P-HQ2F.

[105] *See, e.g.*, *Amazon and Microsoft Win Summary Judgment in Illinois BIPA Lawsuits Based on Extraterritoriality*, JD SUPRA (Nov. 2, 2022), https://perma.cc/E6CU-RDT8.

[106] *See* Steve Lohr, *IBM Tries to Ease Customers' Qualms About Using Generative A.I.*, NY TIMES (Sep. 28, 2023), https://perma.cc/59MS-NKNN; Brad Smith & Hossein Nowbar, *Microsoft Announces New Copilot Copyright Commitment for Customers*, MICROSOFT (Sep. 7, 2023), https://perma.cc/J4DA-2PCV; Emilia David, *Google Promises to Take the Legal Heat in Users' AI Copyright Lawsuits*, THE VERGE (Oct. 12, 2023), https://perma.cc/G98Z-ZC3R.

V. RECOMMENDATIONS

Addressing these issues of fairness and privacy for generative AI technologies is extremely difficult and will require significant resources and attention. A major source of both problems is the foundation of these models: the immense amounts of training data used to develop them. More ethical data collection practices, however, would require a fundamental re-thinking of how data is sourced for AI models and the relationship between developers and the individuals involved in the data creation process. AI has been able to develop and advance very quickly in significant part due to a heavy reliance on web-scraping and other unscrupulous data collection practices, such as using user data for purposes beyond what could have reasonably been foreseen by users.[107] Through such practices, large companies have been able to spend minimal sums to accumulate large amounts of data and instead invest in computing, research, and engineering resources to develop increasingly sophisticated models. This trend is unlikely to stop anytime soon but rather become increasingly problematic as there is a greater push for general-purpose models, whose capabilities rely on increasingly wide-ranging sources of information.

While legal protections alone will not solve these issues given the lack of readily available, large, ethically-sourced datasets, they might be vital for creating incentives for companies to invest in the creation of datasets that are sufficiently diverse, consensual, and non-toxic. Theoretically, if provided sufficient control and assurances regarding how their data would be used and sufficient compensation for its use, many people would be willing to contribute their data for AI development. From personal experience with commissioning bespoke data collection projects, the current market is quite expensive for data that fully complies with relevant privacy and IP laws, is diverse, and ethically sourced, but that does not mean that companies should not make such investments.

This also will require a shift in attention from researchers. Data collection is often under-prioritized in AI research given that academics themselves are accustomed to relying on whatever problematically sourced datasets are publicly available and commonly used,[108] and sourcing data for AI development is often seen as less technically interesting than developing new methods using existing datasets. While the harms of doing academic research using problematic datasets is typically lower than using such datasets to develop products that will be deployed in the real world, the short cycles between research and deployment for cutting-edge AI and the lack of transparency around training data used for commercial AI development underlines the importance of raising ethical standards across the field.[109]

[107] *See, e.g.*, Meta uses users' public Instagram and Facebook posts for training its generative AI models. Mike Clark, *Privacy Matters: Meta's Generative AI Features*, META (Sep. 27, 2023), https://perma.cc/RWR8-BXWN.

[108] *See* Richard Van Noorden, *The Ethical Questions that Haunt Facial-Recognition Research*, 587 NATURE 354, 355-56 (Nov. 18, 2020).

[109] *See id.*

In order to incentivize better data collection practices, an important first step is greater data transparency. Although model cards[110] and data sheets[111] have been commonly cited methods for facilitating such transparency, their adoption is currently entirely voluntary. The transparency obligations for high-risk AI systems under the EU AI Act might provide a useful starting point for creating incentives to improve practices. For foundation models, the EU Parliament's amended version of the Act requires providers to "make publicly available a sufficiently detailed summary of the use of training data protected under copyright law."[112] Even just knowing which datasets are being used for AI development could play a significant role in enabling litigation under existing privacy, IP, and child protection laws.[113] In addition, the prohibition against the "indiscriminate and untargeted scraping of biometric data from social media or CCTV footage to create or expand facial recognition databases,"[114] might further force more consensual data sourcing practices.

In addition to improving data collection practices, more research needs to be dedicated to issues such as bias detection and mitigation, detection of privacy and IP infringement, and model forgetting. Even for companies that want to develop generative AI technologies more ethically, there is currently a dearth of available methods to reliably address these issues at scale. As I have discussed in prior work, there remain many challenges to defining what the appropriate objectives for bias mitigation should be.[115] Should the goal be representative data of the world as it is or should be? If the latter, how do we define what a "fair" world might look like? Generative AI adds further complexities to these questions given that these technologies not only make judgements/predictions but also affirmatively generate content reflecting a specific worldview. There will never be complete consensus on what constitutes "fairness," but any instantiation of AI implicitly adopts a stance, so it is important for developers to consciously and proactively consider how their development decisions might shape this. Regulatory action that requires developers to actively consider bias mitigation can be a beneficial first step. For example, the EU AI Act includes a requirement that providers of foundation models "process and incorporate only datasets that are subject to appropriate data governance

[110] Margaret Mitchell et al., *Model Cards for Model Reporting*, PROC. OF THE ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 220, 220-29 (2019).

[111] Timnit Gebru et al., *Datasheets for Datasets*, PROC. OF THE 5TH WORKSHOP ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY IN MACHINE LEARNING, PMLR 80 (2018), https://perma.cc/8AX8-S7C2.

[112] Amendments Adopted by the European Parliament on 14 June 2023 on the Proposal for a Regulation of the European Parliament and of the Council on Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)), 2024 O.J. (C 506) 1, 173, https://perma.cc/VA45-WNKL [hereinafter EU AI Act Amendments].

[113] Notably researchers recently found that LAION-5B, a popularly used image training dataset that was used by Stability AI in the creation of Stable Diffusion, contained at least 1,679 illegal images featuring CSAM. Emilia David, *AI Image Training Dataset Found to Include Child Sexual Abuse Imagery*, THE VERGE (Dec. 20, 2023), https://perma.cc/G6WB-2FHL.

[114] EU AI Act Amendments, *supra* note 112, at 33 (Amendment 52).

[115] *See* Xiang, *supra* note 29.

measures for foundation models, in particular measures to examine the suitability of the data sources and possible biases and appropriate mitigation." While the focus only on data is limiting, this requirement can still be a valuable first step for forcing the active consideration of problems of bias. Future regulations requiring companies to conduct and share bias evaluations of their models could further be beneficial for ensuring that bias is actually being checked for.

Companies must also invest in internal AI governance structures to provide appropriate accountability for model harms. Especially since it will likely be quite some time until government agencies have the know-how and resources to be able to thoroughly audit increasingly complex and fast-moving AI models, internal red-teaming and research teams dedicated to addressing ethical AI issues will be key. This is especially important given that the real-world harms from AI systems are often difficult to anticipate in a vacuum, without details about the stakeholders and deployment context. For example, there is a big difference between the risk level and mitigation strategies for a generative AI model being used to draft marketing emails versus one being used in an interactive game for children (e.g., a talking character or avatar). In the former case, mitigation strategies would include ensuring there is a human professional in the loop who will check the emails before they are sent out and take responsibility for their content. Training such individuals on possible biases of the model can further reduce the risks. In the latter case, the output of the model will likely need to be highly constrained to avoid it generating problematic content since there will be no human in the loop to check the content before it is shown to end users, who in this case are vulnerable individuals. These types of risk mitigation plans require bespoke analysis that take into consideration the capabilities and safeguards built into the model, the individuals who will be exposed to the AI-generated content, and the possible intervention points in the deployment context, requiring companies to leverage internal assessment processes.

In conclusion, generative AI technologies have been met with immense excitement and anxiety about their potential to revolutionize the capabilities of modern technology and have wide-ranging economic impact. Among the many relevant legal considerations with their proliferation are bias mitigation and privacy protection. In this Essay, I mapped out the potential discriminatory and privacy-related harms that such technologies might present and discussed what we know so far about the legal protections against such harms. In particular, there are lingering questions about the extent to which companies will be incentivized to prevent representational harms from algorithmic bias and to improve their sourcing practices for training data. Addressing these issues through regulatory incentives, along with additional investment and research into bias mitigation methods and more ethical data sourcing practices will be key. Without appropriate incentives in place to address such concerns, we risk living in a world where individuals' control over their personal information is increasingly eroded and existing societal biases are amplified exponentially through the proliferation of and growing reliance on AI-generated content.