Mirror, Mirror, on the Wall, Who's the Fairest of Them All?

Alice Xiang

Debates in AI ethics often hinge on comparisons between AI and humans: which is more beneficial, which is more harmful, which is more biased, the human or the machine? These questions, however, are a red herring. They ignore what is most interesting and important about AI ethics: AI is a mirror. If a person standing in front of a mirror asked you, "Who is more beautiful, me or the person in the mirror?" the question would seem ridiculous. Sure, depending on the angle, lighting, and personal preferences of the beholder, the person or their reflection might appear more beautiful, but the question is moot. AI reflects patterns in our society, just and unjust, and the worldviews of its human creators, fair or biased. The question then is not which is fairer, the human or the machine, but what can we learn from this reflection of our society and how can we make AI fairer? This essay discusses the challenges to developing fairer AI, and how they stem from this reflective property.

How can we develop fairer artificial intelligence (AI) that does not reflect, entrench, and amplify societal biases? There are three major categories of interventions: data curation, algorithmic methods, and policies around appropriate use. The first is motivated by the fact that AI, like a mirror, tends to reflect the biased patterns present in its training data. If a voice recognition model is trained predominantly on audiobooks, it might learn how to accurately understand "standard" varieties of language but struggle to understand accents, dialects, or speech impediments.[1] In domains like computer vision and speech language technologies, diversity in the appearance and voices of the individuals represented in the training data is key to avoid the creation of biased AI models.[2] The second set of approaches is algorithmic interventions. AI is not a perfect mirror, so by imposing constraints or changing the objectives for the model's optimization, algorithmic fairness practitioners seek to warp the mirror, making the outputs more accurate or fairer. Much of the literature in this space focuses on defining specific fairness metrics and developing preprocessing, in-processing, or postprocessing methods to make the model's outputs perform better on the basis of those fairness metrics.[3] The third set of approaches focuses on defining when AI or humans should be used. For example, the moratoriums several U.S.

https://doi.org/10.1162/daed_a_02058

jurisdictions put in place around law enforcement's use of facial recognition and the European Union's AI Act, which prohibits certain high-risk categories of AI, fall under this category.[4] Combinations of these approaches are vital to addressing bias mitigation, but as this essay will discuss, there are many technical, legal, and operational challenges to creating fairer AI in practice.

Starting first with data curation, regarding AI as a mirror implies looking beyond the AI model itself to the societal context surrounding it, which it reflects in turn. Just as a parent can shape their child's worldview by controlling the information and experiences the child is exposed to, AI developers can similarly mold their AI through their data selection. Most image datasets are sourced exclusively from a few developed countries.[5] Biases in computer vision models have largely been attributed to a lack of sufficient representation of women and minorities in such datasets.[6] Like humans who find it easier to accurately distinguish people in the majority ethnic group they grew up in, human-centric computer vision models tend to more accurately recognize the types of people featured in their training data.[7] Moreover, lack of diversity in the background, objects, clothing, and other features can lead to additional biases. For example, what does "soap" look like? The answer can differ depending on which part of the world you are from. Researchers found that object detection algorithms trained predominantly on data from higher-income countries struggled to accurately recognize objects in lower-income countries.[8]

The digital divide can further exacerbate inequities in whose interests are reflected in datasets. For example, in 2012, Boston-based startup Connected Bits launched its StreetBump app, leveraging accelerometer and GPS data to automatically detect potholes and inform the city where to direct resources to fix them.[9] The data collected by the app, however, painted a distorted picture of the prevalence of potholes in the city.[10] People in lower-income neighborhoods were less likely to have smartphones to download the app, leading to systematic underrepresentation of the number of potholes in their neighborhoods needing repair.

There are thus strong normative arguments for collecting carefully curated, large, diverse, representative datasets to tackle algorithmic bias. While this statement has become a truism in the algorithmic fairness community, how to collect such datasets in practice is an unsolved problem.[11] Much of the progress in the past few decades of AI development has stemmed directly from questionable data-collection practices. In the early days of computer vision, images were sourced in highly controlled bespoke settings.[12] Researchers would set up photography studios to take pictures of subjects. These image datasets were consequently very small and highly constrained. The poses, backgrounds, and demographics of the people represented in these datasets were greatly limited. Computer vision and AI more generally were revolutionized by the development of large, publicly available web-scraped datasets. ImageNet, consisting of fourteen million images

scraped from the internet with annotations for the objects in the images, was revolutionary in enabling computer vision scientists to train their models on much larger-scale data than was previously possible.[13] In the ImageNet Large Scale Visual Recognition Challenge, AlexNet (a convolutional neural network) won, substantially beating the runner-up. A key feature of AlexNet was the depth of its network, which relied on a large training dataset. The success of AlexNet, one of the most influential developments in computer vision, contributed to the explosion in deep learning.[14]

While ImageNet and AlexNet were tremendously beneficial for the acceleration of AI development, they also set AI developers along a path that depended on vast amounts of data and computation. Achieving state-of-the-art models required large corpora of data that could not easily be obtained through curated, bespoke methods. Web-scraping, the method by which ImageNet was created, became the norm, with most large datasets since ImageNet relying on that method.[15] While web-scraping large amounts of online data leveled the playing field to some extent, it also carried with it significant ethical challenges. In recent years, ImageNet has faced criticism for its lack of informed consent, offensive labels, and problematic images, all of which are artifacts of its collection methodology.[16]

This dependency on web-scraped images has carried over to algorithmic fairness efforts. My recent work has discussed this issue in depth, exploring the tensions that emerge between fairness and privacy in operationalizing data-collection efforts for human-centric computer vision.[17] For example, in 2018, IBM released the Diversity in Faces (DiF) dataset.[18] Like most large-scale computer vision datasets at the time, this dataset was based on images scraped from Flickr with permissive licenses. IBM's contribution was to find a diverse subset of face images and provide labels of relevant features, enabling the dataset to be used by fairness researchers checking for biases in their models. Even though ImageNet, COCO (Common Objects in Context, another large web-scraped image dataset), and other major datasets similarly featuring Flickr images with humans had been available for years without any lawsuits, the launch of DiF was immediately fraught. Not only was IBM sued under the Illinois Biometric Information Privacy Act (BIPA) for processing individuals' biometric information without appropriate informed consent, but Microsoft and Google were also sued as downstream users of the dataset.[19] DiF was immediately removed by IBM, and not long afterward, IBM announced that it would be pulling away from facial recognition technologies in general.[20] Notably, other Flickr-based computer vision datasets remain available and have not faced any lawsuits. In 2021, ImageNet creators voluntarily decided to obscure the faces of image subjects (note that their bodies are not otherwise obscured), but COCO remains available without any obfuscation of faces or bodies.[21] Taking the DiF dataset as a starting point, let us consider the minefield of constructing a "fair" human image dataset.

For simplicity, I will consider "fair" to simply mean a dataset that is legally compliant, as globally diverse and free of biases as possible, and large and realistic enough to develop a state-of-the-art model. The benefits of web-scraped datasets are that they are large and realistic. That is not to say that they are free from biases. In fact, they tend to exhibit biases reflective of the platform aggregating the data and of society as a whole. For example, studies have shown that images on Flickr tend to be biased toward Western developed countries, where most of Flickr's users are located.[22] In addition, AI trained on such datasets tend to learn stereotypical patterns, such as associating women with domestic spheres and men with public spheres.[23] For instance, commonly used visual recognition datasets feature women cooking far more often than men cooking, teaching AI models to associate women with the activity of cooking.[24] Similar stereotypical trends have been found in word embeddings from large text corpora that can be used to train language models.[25]

Moreover, as the DiF authors discovered, web-scraping presents many legal issues. In the past few years, lawsuits stemming from U.S. state biometric information privacy laws and the European Union's General Data Protection Regulation have raised awareness that using face images for AI without informed consent is inappropriate from a legal compliance perspective. Facebook reached a landmark settlement in 2021 for $650 million in a BIPA lawsuit contesting their processing of users' biometric data through facial recognition technology to support their automated tag-suggestion feature.[26] Nonetheless, researchers in computer vision still rely heavily on such datasets. Many do not see any alternative for how they could conduct research in this field otherwise.[27] Indeed, the recent explosion in generative AI technologies has only further exacerbated this issue, requiring even larger amounts of data to train, and normalizing the idea that participation at the frontier of AI necessitates training such models indiscriminately on content from across the internet.

Copyright has also increasingly become a concern in data curation. For models trained on large web-scraped corpora of text, images, and video, rarely is the permission of content creators sought prior to using their content for AI development. This has especially presented problems in the generative AI context, where AI models not only benefit from the use of copyrighted content but often generate new content inspired by such inputs, but without appropriate attribution. Creators of web-scraped computer vision datasets have historically emphasized their reliance on data with permissive licensing as an argument for why they are not infringing on IP rights.[28] While such arguments might have sufficed when copyrighted materials were used to train AI models for tasks unrelated to creative pursuits – such as transcribing text or drawing bounding boxes, key points, or segmentation masks on humans and objects – generative AI presents new concerns. If an artist uploads their work to a public platform with a permissive license for the content to be redistributed, have they

also agreed to allow the generation of derivative works that might imitate their style or content, possibly to the extent of cannibalizing their business? Arguably, until recently, few artists could have foreseen such consequences.

In addition, collecting globally representative data presents many practical complexities due to real-world geopolitical divisions. Contracting with and obtaining informed consent from people around the world is challenging given differences in local laws, cultures, and languages. Privacy and intellectual property rights vary substantially across jurisdictions. Many countries also have data localization laws that erect barriers to the transfer of their residents' data outside of their country.[29] Economic sanctions can further affect the extent to which some countries can be represented in AI datasets. These constraints add a geopolitical dynamic to what AI models learn about the world. Similar to how China's "Great Firewall" has led to a distinct internet experience for Chinese netizens, legal and political barriers around data collection can lead to more fragmented AI development, with parochial AI models that primarily only understand the people and patterns in their own geographies.[30]

Beyond web-scraping, another approach to assembling large, diverse datasets is to use existing repositories of stock images taken by professional photographers. While this is less problematic than the web-scraping approach given that photographers and the image subjects were possibly compensated for their work, it is difficult to say whether these individuals could have anticipated that their works would be used to develop AI. Especially in an era of generative AI, allowing your photos to be used to train AI could have downstream implications, such as content featuring your likeness (if you are the image subject) or artistic style (if you are the photographer) being generated by the AI in response to prompts you have no control over. This is especially problematic if the generated content is misleading or offensive. Moreover, stock photos taken by professional photographers look very different from the more naturalistic images that AI is likely to encounter in deployment. The lighting is often perfected, the setting and poses staged, and the image subjects more conventionally attractive. AI developed using such images can have a harder time recognizing people or objects in the real world due to this domain shift, or difference in the distributions of the training data and deployment context data.[31]

Bespoke data collection that reflects the deployment context is thus generally the best approach in that it enables more control over the data-collection process and assurance that both artists and subjects are fully informed about how their data are likely to be used. Operationalizing bespoke data collection, however, is very difficult. It requires developing business relationships with people around the world who can contribute to data-collection efforts. As a result, many companies specialize specifically in data collection. They recruit large numbers of crowd-workers from around the world to perform various data collection and an-

notation tasks for client companies. These companies have faced significant scrutiny, however, as some have deployed problematic recruitment practices to source diverse crowds and others have failed to provide appropriate employment conditions for data annotators.[32] Collecting billions of images through such methods can also be cost-prohibitive for smaller companies and researchers.[33]

Bespoke data collection further presents the challenge of requiring diversity specifications. In an underspecified dataset for which demographic balance is required of only a few attributes, like gender, age, and ethnicity, the images tend to look highly staged and homogenous.[34] It is easiest for people to take pictures of themselves standing or sitting, facing the camera, inside or right outside of their home. Additional requirements, such as a variety of poses, backgrounds, lighting conditions, number of people/objects, and interactions between them, can exponentially increase the complexities of data collection. This is especially the case given that it is not enough to provide a generic specification that diversity along these dimensions should be maximized. Checking for and ensuring diversity requires annotations specifying what pose, background, and lighting conditions are featured in the image. This requires a taxonomy for such attributes and extensive time and resourcing for image subjects or annotators to label the images. How do you adequately define the parameters for how the "real world" looks?

Moreover, the annotations related to the diversity of image subjects themselves can be highly contentious. For example, there are often concerns around bias associated with race, ancestry, or ethnicity, but collecting data on these attributes to check for bias can be complex given the social construction of such attributes. Different countries vary widely in how their census surveys characterize relevant ethnic groups, with some even refusing to collect race data, so there is no singular taxonomy that is consistent across the world.[35] Even the act of asking someone for these attributes can raise privacy concerns given the sensitive nature of such data, along with worries that the data will be used to discriminate against them (rather than prevent discrimination).[36]

Given these challenges with real data, there has been growing interest in the potential for synthetic data, leveraging recent advances in generative AI. Bypassing the need for real people, synthetic data can reduce many of the legal challenges with using real data, but issues of fairness persist. Creating synthetic data is like creating a microcosm of the world: while developers might be freed from some of the constraints of reality, that freedom also creates more room for subjectivity. For example, every conceivable skin tone, nose/face/eye shape, hairstyle, or body type is theoretically possible to generate synthetically, but with this flexibility comes more need for developers to specify the parameters of interest. It is like asking an artist to draw a fully inclusive representation of humankind. Biases and limitations of the artist's imagination can translate into a narrower worldview compared with large amounts of real-world data. For example, an artist might draw figures of vary-

ing skin tones and facial features, but all with similar body types and clothing styles, with backgrounds and objects that reflect a middle-class American living standard. Like a parent raising their child in a virtual simulation, AI developers who rely on synthetic data theoretically have more control over what their "child" is exposed to, but it can be difficult to create a synthetic environment as rich as reality but lacking the biases of the real world. For AI that operates exclusively in a synthetic environment, like AI avatars in video games, such a domain shift is not necessarily a problem. In most cases where the AI interacts with the real world, however, algorithmic bias is relevant, and this difference between the "world" where the AI is developed versus deployed can exacerbate potential biases.

Addressing data diversity and sourcing, however, is only the first part of the problem. Having a globally representative dataset simply ensures that the mirror is not warped, and your model reflects a more accurate representation of the world. The reflection we see in a perfect mirror is nonetheless often not flattering. Societal inequities and injustices that are present in the real world will naturally be reflected in such data. This presents one of the major challenges of algorithmic fairness: how to conceptualize a fair society and enable our AI models to promote rather than work against such a conception.

Early work highlighted the challenge of optimizing for multiple fairness definitions simultaneously. Researchers quickly proved impossibility theorems showing that some of the common fairness metrics conflicted with each other. Specifically, a model could not simultaneously be well-calibrated and have equalized odds across demographic groups if the demographic groups had different baselines.[37] The impossibility theorem inspired greater technical interest in the problem of algorithmic fairness.[38]

While the idea that data might reflect problematic patterns is increasingly accepted, the question of how to address these patterns is much less clear. While the algorithmic fairness literature features many solutions that imply differing thresholds or quotas for various sensitive attribute groups (that is, attributes receiving special legal protections like race, gender, or age), such solutions could be highly suspect viewed through a legal lens. As scholars have recently highlighted, it might not seem immediately evident that Supreme Court deliberations on affirmative action in higher education might have any bearing on algorithmic fairness.[39] But there are strong parallels that imply that if there were federal antidiscrimination litigation around algorithmic bias mitigation, many of the proposed methods could be deemed illegal.

In recent decades, the Supreme Court has increasingly turned toward anticlassification doctrine in its rulings.[40] Anticlassification is akin to colorblindness and implies that the fundamental goal of antidiscrimination law should be to prevent differential classification or treatment of individuals based on their protected

attributes. This contrasts with antisubordination, the doctrine that holds that antidiscrimination law should seek to actively dismantle historical discriminatory structures. Lyndon Johnson famously articulated the antisubordination underpinnings of affirmative action during his 1965 commencement address at Howard University: "You do not take a man who for years has been hobbled by chains, liberate him, bring him to the starting line of a race, saying, 'You are free to compete with all the others,' and still justly believe you have been completely fair."[41]

While debates about affirmative action have been active and controversial for many decades, the algorithmic fairness context highlights unique dimensions.[42] Cases like *Bakke*, *Grutter*, and *Gratz* conveyed the message to schools that affirmative action is only permissible if it cannot be easily quantified.[43] Quotas and point systems were patently unconstitutional, whereas holistic systems that used race as one of many factors were permissible. These types of decisions provided actionable guidance for human admissions officers who could keep an eye on racial composition of the class when making decisions, without ever formally quantifying any affirmative action boost. Such obfuscation is much more difficult for an algorithm.[44]

But the recent *Students for Fair Admissions* joint decision closed off even these approaches, solidifying the court's adoption of the anticlassification stance, as it struck down the affirmative action programs at Harvard and the University of North Carolina.[45] On the one hand, the court faulted these universities for their failure to provide quantifiable metrics for success (such as how much diversity is sufficient to obtain their educational objectives). But on the other hand, the court found their programs to be unconstitutional for the implicit quotas they adopted: for Harvard, "how the breakdown of the class compares to the prior year in terms of racial identities," and for the University of North Carolina, whether the "percentage enrollment within the undergraduate student body is lower than their percentage within the general population in North Carolina."[46] The court also declared that race could never be used as a negative factor, which in the zero-sum game of college admissions, implied that race could not be considered directly as a factor.[47] The only allowance the court gave to schools was that they could consider based on applicants' essays the possible impact of race on their experiences, provided that such essays highlighted the applicants' courage, determination, or other positive attributes.[48] The court has thus left very little room for explicit race-conscious antidiscrimination interventions, potentially posing challenges for the algorithmic fairness community, whose work typically involves formalizing a fairness metric, constraint, or objective that is conscious of the protected attribute, with the goal of affirmatively changing the model to be "fairer."[49]

The technical formalism of AI ethics, however, can also be used to reframe these contentious societal debates with greater clarity. At a time when it is common to bemoan the bias, opacity, and lack of accountability of AI, which is increasingly used throughout our society, does it make sense to incentivize either ignorance or

obfuscation of biases in such technology? AI developers seeking technologies that do not perpetuate societal biases already encounter many challenges to even testing for bias. As discussed above, privacy laws strongly disincentivize the collection of sensitive attribute data that is necessary for conducting bias audits. Should legal doctrine in antidiscrimination law further disincentivize developers from taking any action once bias has been discovered, out of fear of being successfully sued for (reverse) discrimination?

Understanding the connection between algorithmic fairness and broader societal debates about equity thus raises the stakes of these debates. Not only are courts debating the admissions criteria for elite schools, but such legal decisions codify normative principles that can influence the extent to which developers are legally allowed to modify increasingly ubiquitous algorithms to avoid amplifying bias against people from marginalized communities, despite their sway over decisions around recidivism, employment, credit, or other high-stakes domains. In other words, there may be a limit to how much developers can do to reduce the harm done by their own work.

At the heart of such societal debates is the tension between erasing versus mitigating the effects of systemic discrimination. Outside of the algorithmic context, proponents of anticlassification would argue that the goal should be to desensitize people to sensitive attributes like race and pursue a colorblind society.[50] Algorithmic fairness questions this notion given that AI trained without features like race are by default colorblind yet can still be racist. The richness of big data implies the presence of proxy variables and patterns correlated with race and other sensitive attributes that can be learned by a model that is not explicitly given sensitive attribute data.[51] For example, in the famous COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) case, the algorithm did not have any direct information about the race of the defendants. Nevertheless, because the training data reflected broader national trends whereby Black defendants had higher rates of re-arrest, likely due to disproportionate policing practices, the COMPAS algorithm leveraged features correlated to race, such that Black individuals were more likely to be incorrectly labeled as having high-recidivism risk.[52] Unlearning or avoiding biases on an algorithmic level typically requires knowledge of sensitive attribute information.[53] Algorithmic fairness thus errs on the side of antisubordination. While humans might be able to argue that ignoring race is an effective way to address racism, this is much more difficult for AI. Teaching AI the notion that racism is bad and should be avoided requires providing models some data about race.

Another way to view this debate through the lens of algorithmic fairness is to consider statistician George Box's famous quote, "All models are wrong, but some are useful."[54] Although he was addressing statistical models more generally rather than AI models, let alone bias in AI models, his insights still ap-

ply. Opponents of algorithmic bias mitigation efforts often resist interventions that are motivated by social justice inclinations out of concern that they are tampering with what is correct, true, or accurate. Indeed, the fact that fairness and accuracy in AI are often framed in terms of a trade-off is reflective of this idea.[55] The reality, however, is that all AI models are approximations of reality as conveyed to them via the data they are trained on. They are approximations built upon approximations of reality, and thus riddled with inaccuracies. For example, researchers found bias in health care algorithms that used cost of care as a proxy for health care need.[56] The training data reflected the pattern that Black patients of similar sickness levels to white patients receive less health care, so the model learned to downgrade the risk level of Black patients. Acknowledging these imperfections, the question then is how should we correct them? Bias mitigation efforts, instead of being framed as introducing additional inaccuracies, should be viewed as correcting existing inaccuracies in a direction that is more favorable from an equity perspective.[57]

This distinction can also be framed as a difference between prediction versus decision-making. Is the goal to have a mirror that as accurately as possible reflects reality in order to make accurate predictions? Or is the goal to improve upon the world and make it fairer? If AI is used purely for predictive tasks, like predicting whether someone will be re-arrested, then bias mitigation is less relevant given that reflecting societal biases accurately is helpful for making accurate predictions. But if AI is used for decision-making, there is a normative element that implies a need for bias mitigation. For example, deciding who should be denied bail is different from predicting who might be re-arrested. Many harms related to AI ethics stem from the conflation of a normative task with a descriptive one. If the goal is to decide who should be detained because they are more likely to commit a crime, then it is important to separate the bias of over-policing from the ground truth of crimes committed. This separation process is precisely what bias mitigation should aim to do.

The rise of generative AI technologies might further bring these debates into the content generation sphere. What content should be considered biased or discriminatory? These questions have long challenged content platforms, which typically rely on a combination of community guidelines, automated flagging of objectionable content, user reports, and human content moderators. Such efforts have thrown content platforms into contentious societal debates around whether their efforts are reasonable corrections to avoid disinformation versus problematic distortions of free speech.[58] With generated content from AI, however, the debate shifts: the question is no longer what human content is permissible to be shared on a platform, but rather what AI content should be generated. If an image generator consistently generates images of men whenever prompted with terms like "CEO," "intellectual," or "director," the AI might entrench existing societal

stereotypes.[59] To the extent such AI-generated images are then used to make or inspire art, movies, or media, they will amplify these biases.

On the one hand, from an antisubordination standpoint, this should create a responsibility on the part of the AI developer to take active measures to ensure the content generated reflects a less-biased view of the world. On the other hand, given current controversies around content moderation policies, it is likely that such affirmative efforts to create more balanced representations will be politically fraught.

In light of all of these challenges to implementing bias mitigation in practice, it is worth addressing the skepticism as to whether such efforts should even be pursued. Any fairness efforts predicated on having access to diverse data or sensitive attribute data necessitates the collection of yet more data, often about people from vulnerable, underrepresented populations, creating potential tradeoffs between fairness and other values like privacy.[60] Any attempts to rebalance the benefits or harms of algorithmic systems across demographic groups might cause significant political controversy, as the previous section discussed. It is tempting for such debates to go to extremes – for example, concluding that privacy must be protected at all costs – so fairness efforts requiring the collection of sensitive information at scale should be immediately halted.

For instance, some scholars have highlighted the concept of "horizontal relationality" in privacy, whereby the disclosure of private information by one individual could affect another individual's privacy, particularly in the context of machine learning and AI.[61] An example they use is if someone shares an image of their tattoo for a tattoo recognition model, the inclusion of their tattoo image in the training data for the tattoo recognition model could affect the model's ability to recognize similar tattoos on other individuals. If the tattoo recognition model is used by law enforcement to identify potential suspects, this could impact those individuals' privacy.

While horizontal relationality has been primarily characterized in a negative light – how one person's sacrifice of their privacy can force others to sacrifice their privacy – what's lost in this discussion is that there are benefits to horizontal relationality as well. In particular, such analyses assume an antagonistic relationship with technology, in which the goal is to ensure that AI does not work well for you. Such an attitude is typically motivated by concerns around AI surveillance: that AI primarily is being deployed by governments, employers, and others with power to surveil and deprive lower-powered individuals of autonomy and self-determination.[62] Many AI applications, however, lack this antagonistic relationship. Individuals buying a camera with AI autofocus for use in taking personal photos generally want the camera to be able to focus on their faces, their family's faces, and their friends' faces as accurately as possible. There is not necessarily a surveillance risk if the individual is taking photos and storing them on their drive

for personal consumption. While theoretically the person sharing images publicly on social media might create some surveillance risk for the individuals in the photos, that is unrelated to the functionality of the AI autofocus. If the autofocus worked poorly, the individual would likely just spend more time trying to get a good shot rather than give up entirely on sharing their lives on social media.

Even in high-stakes scenarios like law enforcement use of facial recognition to find suspects, it is unclear that reducing the performance of the AI model provides any benefits from the perspective of reducing surveillance. Much of the outcry against such high-risk use cases stem precisely from the negative impacts of poor performance of such models. In particular, there have been several notable cases in the United States of Black men being wrongfully arrested due to faulty facial recognition matches.[63] The question of whether law enforcement use of facial recognition is acceptable (a topic beyond the scope of this essay) is distinct from the question of whether better or worse accuracy of technologies is preferable.

If we assume such technologies will continue to be in use, then better accuracy benefits everyone other than those trying to evade law enforcement. Misrecognition for individuals with less societal privilege is especially pernicious since these individuals are less likely to have access to recourse to prove the mistake. This could include access to effective legal counsel, knowledge of relevant legal protections, and funds needed for bail. But weakening such surveillance technologies or making them less accurate won't benefit those people most harmed now; it will simply make such wrongful arrests *more* likely. So, if such surveillance technologies are in widespread use, there is a strong argument for maximizing the accuracy of such technologies for all groups, with the greatest benefits to those who are most victimized by the errors. More generally, regardless of whether a technology is low or high risk, trying to combat biases in the technology is a worthy endeavor. Critiques about algorithmic fairness efforts would more accurately be framed as critiques of specific AI use cases, especially ones that are surveillance oriented.

Separating these two considerations – whether a technology should be banned versus whether it should be improved – is of critical importance when attempting to operationalize algorithmic fairness. Why should humanity not have its cake and eat it too? While there might be some technologies so dangerous that outright bans are the only morally permissible response, the majority of AI technologies fall into a gray area in which their use should be conditional on appropriate safeguards. Navigating such gray areas requires taking action to address issues of bias, even when doing so requires carefully balancing other ethical desiderata.

Compared to other forms of technology, a distinguishing feature of AI is its capacity to learn from the data presented to it. This learning process transforms AI from a purely objective, rational machine to a mirror reflecting a version of our world. What makes AI ethics a fascinating discipline is that the

problems in this subfield are a microcosm for broader societal problems. The key difference, however, is that AI is our own creation, which sets a stronger moral requirement for us to address these problems and avoid employing AI that perpetuates and entrenches existing societal problems. Moreover, in certain ways, we have more control over AI models than we do over broader society. For example, although collecting a globally diverse training dataset to train a facial recognition model is extremely difficult, it is still easier than counteracting the biases of billions of peoples' human facial recognition. Thus, developing fairer AI is a difficult task, not simply because AI is often a black box, but also because AI reflects society and all its complexities. AI developers are often faced with difficult unsolved ethical questions that cut to the core of contemporary debates: What should you do to rectify historical injustices? How can you achieve fairness or diversity? To address these questions, we must think of AI not as a separate entity, a jumble of numbers and code, but rather as a mirror reflecting our society.

ABOUT THE AUTHOR

Alice Xiang is Global Head of AI Ethics at Sony Group Corporation and Lead Research Scientist at Sony AI. She has published in such journals as the *Harvard Journal of Law & Technology*, *Tennessee Law Review*, *University of Chicago Law Review*, and *Yale Law Journal* and in publications from machine learning conferences like *NeurIPS*, *ICML*, *ICLR*, *ICCV*, and *FAccT*.

ENDNOTES

[1] Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, Jerone Andrews, et al., "Augmented Datasheets for Speech Datasets and Ethical Decision-Making," paper presented at the ACM Conference on Fairness, Accountability & Transparency, Chicago, Ill., June 12, 2023, https://arxiv.org/abs/2305.04672.

[2] Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research* 81 (2018): 1–15, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[3] B. d'Alessandro, Cathy O'Neil, and Tom LaGatta, "Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification," *Big Data* 5 (2) (2017): 120–134.

[4] Electronic Privacy Information Center, "State Facial Recognition Policy," https://epic.org/state-policy/facialrecognition (accessed December 12, 2023); and European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* (Luxembourg: Publications Office of the European Union, 2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[5] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten, "Does Object Recognition Work for Everyone?" paper presented at the Institute of Electrical and Electronics Engineers/Computer Vision Foundation Conference on Computer Vision and Pattern Recognition, Workshop 52, Long Beach, Calif., June 16, 2019, https://openaccess.thecvf.com/content_CVPRW_2019/html/cv4gc/de_Vries_Does_Object_Recognition_Work_for_Everyone_CVPRW_2019_paper.html; and Keziah Naggita, Julienne LaChance, and Alice Xiang, "Flickr Africa: Examining Geo-Diversity in Large-Scale, Humancentric Visual Data," paper presented at the Association for the Advancement of Artificial Intelligence/Association for Computing Machinery Conference on AI, Ethics and Society, Montreal, August 8, 2023.

[6] Buolamwini and Gebru, "Gender Shades."

[7] P. Jonathon Phillips, Fang Jiang, Abhijit Narvekar, et al., "An Other-Race Effect for Face Recognition Algorithms," *ACM Transactions on Applied Perception* 8 (2) (2011): 1–11, https://doi.org/10.1145/1870076.1870082; and Elinor McKone, Lulu Wan, Madeleine Pidcock, et al., "A Critical Period for Faces: Other-Race Face Recognition Is Improved by Childhood But Not Adult Social Contact, Scientific Reports," *Nature*: *Scientific Reports* 9 (2019), https://www.nature.com/articles/s41598-019-49202-0.

[8] DeVries, Misra, Wang, et al., "Does Object Recognition Work for Everyone?"

[9] Connected Bits, "Street Bump," https://connectedbits.com/street_bump (accessed December 12, 2023).

[10] Kate Crawford, "The Hidden Biases in Big Data," *Harvard Business Review*, April 1, 2013, https://hbr.org/2013/04/the-hidden-biases-in-big-data.

[11] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, et al., "Advances, Challenges, and Opportunities in Creating Data for Trustworthy AI," *Nature Machine Intelligence* 4 (2022): 669–677, https://doi.org/10.1038/s42256-022-00516-1; and Alice Xiang, "Being 'Seen' vs. 'Mis-Seen': Tensions between Privacy and Fairness in Computer Vision," *Harvard Journal of Law & Technology* 36 (1) (2022): 1–60, https://doi.org/10.2139/ssrn.4068921.

[12] Inioluwa Deborah Raji and Genevieve Fried, "About Face: A Survey of Facial Recognition Evaluation," paper presented at the Association for the Advancement of Artificial Intelligence 2020 Workshop on AI Evaluation, New York, February 7, 2020, https://doi.org/10.48550/arXiv.2102.00813.

[13] ImageNet, https://www.image-net.org (accessed December 12, 2023).

[14] Li Fei-Fei and Ranjay Krishna, "Searching for Computer Vision North Stars," *Dædalus* 151 (2) (Spring 2022): 85–99, https://www.amacad.org/publication/searching-computer-vision-north-stars.

[15] Raji and Fried, "About Face."

[16] Vinay Uday Prabhu and Abeba Birhane, "Large Image Datasets: A Pyrrhic Win for Computer Vision?" paper presented at the Institute of Electrical and Electronics Engineers/Computer Vision Foundation Conference on Applications of Computer Vision, 2021, https://openaccess.thecvf.com/content/WACV2021/papers/Birhane_Large_Image_Datasets_A_Pyrrhic_Win_for_Computer_Vision_WACV_2021_paper.pdf.

[17] Xiang, "Being 'Seen' vs. 'Mis-Seen.'"

[18] Michele Merler, Nalini Ratha, Rogerio Feris, and John R. Smith, "Diversity in Faces," arXiv, April 10, 2019, https://arxiv.org/pdf/1901.10436.pdf.

[19] Taylor Hatmaker, "Lawsuits Allege Microsoft, Amazon and Google Violated Illinois Facial Recognition Privacy Law," Tech Crunch, July 15, 2020, https://techcrunch.com/2020/07/15/facial-recognition-lawsuit-vance-janecyk-bipa.

[20] Nicolas Rivero, "The Influential Project that Sparked the End of IBM's Facial Recognition Program," *Quartz*, June 10, 2020, https://qz.com/1866848/why-ibm-abandoned-its-facial-recognition-program.

[21] Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, et al., "A Study of Face Obfuscation in ImageNet," arXiv, March 10, 2021, last revised June 9, 2022, https://doi.org/10.48550/arXiv.2103.06191; Common Objects in Context, https://cocodataset.org (accessed December 12, 2023); and Julienne LaChance, William Thong, Shruti Nagpal, and Alice Xiang, "A Case Study in Fairness Evaluation: Current Limitations and Challenges for Human Pose Estimation," paper presented at the Association for the Advancement of Artificial Intelligence 2023 Workshop on Representation Learning for Responsible Human-centric AI (R2HCAI), Washington, D.C., February 13, 2023, https://r2hcai.github.io/AAAI-23/files/CameraReadys/21.pdf.

[22] Abhishek Mandal, Susan Leavy, and Suzanne Little, "Dataset Diversity: Measuring and Mitigating Geographical Bias in Image Search and Retrieval," *Proceedings of the First International Workshop on Trustworthy AI for Multimedia Computing* (Trustworthy AI '21): 19–25, https://doi.org/10.1145/3475731.3484956; and Naggita, LaChance, and Xiang, "Flickr Africa."

[23] Angelina Wang, Alexander Liu, Ryan Zhang, et al., "REVISE: A Tool for Measuring & Mitigating Bias in Visual Datasets," paper presented at the European Conference on Computer Vision, virtual, August 23, 2020, https://arxiv.org/pdf/2004.07999.pdf.

[24] Jieyu Zhao et al., "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-Level Constraints," paper presented at the Conference on Empirical Methods in Natural Language Processing, Copenhagen, September 7, 2017, https://arxiv.org/abs/1707.09457; and Dora Zhao, Jerone Andrews, and Alice Xiang, "Men Also Do Laundry: Multi-Attribute Bias Amplification," *Proceedings of Machine Learning Research* 202 (2023): 42000–42017, https://arxiv.org/abs/2210.11924.

[25] Tessa E. S. Charlesworth, Victor Yang, Thomas C. Mann, et al., "Gender Stereotypes in Natural Language: Word Embeddings Show Robust Consistency Across Child and Adult Language Corpora of More Than 65 Million Words," *Psychological Science* (2021): 1–23.

[26] *In re: Facebook Biometric Information Privacy Litigation*, 185 F. Supp. 3d 1155 (N.D. Cal. 2016), Order Re Final Approval, Attorneys' Fees and Costs, and Incentive Awards; Signed by Judge James Donato on February 26, 2021, https://www.govinfo.gov/app/details/USCOURTS-cand-3_15-cv-03747/USCOURTS-cand-3_15-cv-03747-16.

[27] Richard Van Noorden, "The Ethical Questions that Haunt Facial Recognition Research," *Nature* 587 (2020): 354–358, https://doi.org/10.1038/d41586-020-03187-3.

[28] Prabhu and Birhane, "Large Datasets."

[29] Richard D. Taylor, "'Data Localization': The Internet in the Balance," *Telecommunications Policy* 44 (8) (2020), https://doi.org/10.1016/j.telpol.2020.102003.

[30] James Griffiths, "Acronyms and Abbreviations," in *The Great Firewall of China: How to Build and Control an Alternative Version of the Internet* (London: Zed Books Ltd, 2019), xi–xii.

[31] Daniel E. Ho, Emily Black, Maneesh Agrwawala, and Fei-Fei Li, "Domain Shift and Emerging Questions in Facial Recognition Technology," policy brief, Stanford Uni-

versity Human-Centered Artificial Intelligence, https://hai.stanford.edu/sites/default/files/2020-11/HAI_FRT_WhitePaper_PolicyBrief_Nov2020.pdf.

[32] Sidney Fussell, "How an Attempt at Correcting Bias in Tech Goes Wrong," *The Atlantic*, October 29, 2019, https://www.theatlantic.com/technology/archive/2019/10/google-allegedly-used-homeless-train-pixel-phone/599668; and Billy Perrigo, "Inside Facebook's African Sweatshop," *Time*, February 14, 2022, https://time.com/6147458/facebook-africa-content-moderation-employee-treatment.

[33] Xiang, "Being 'Seen' vs. 'Mis-Seen.'"

[34] Ibid.

[35] Marie des Neiges Leonard, "Census and Racial Categorization in France: Invisible Categories and Color-Blind Politics," *Humanity & Society* 38 (1) (2014): 67–88; and Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker, "How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis," *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW1) (2020): 1–35, https://doi.org/10.1145/3392866.

[36] Jessica L. Roberts, "Protecting Privacy to Prevent Discrimination," *William and Mary Law Review* 56 (6) (2015): 2097–2174, https://scholarship.law.wm.edu/wmlr/vol56/iss6/4.

[37] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," paper presented at the Eighth Conference on Innovations in Theoretical Computer Science, Cambridge, Mass., January 14, 2016, https://arxiv.org/abs/1609.05807.

[38] Sam Corbett-Davies, Emma Pierson, Avi Feller, et al., "Algorithmic Decision Making and the Cost of Fairness," paper presented at the 23rd ACM SIGKDD International Conference on Knowledge, Discovery & Data Mining, Halifax, August 13–17, 2017; Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, et al., "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning," *Journal of Machine Learning Research* 24 (2023), https://www.jmlr.org/papers/v24/22-1511.html; Alexandra Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data* 5 (2) (2017): 153–163; and James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan, "An Intersectional Definition of Fairness," paper presented at the 36th Institute of Electrical and Electronics Engineers International Conference on Data Engineering, Dallas, Tex., April 21, 2020.

[39] Alice Xiang, "Reconciling Legal and Technical Approaches to Algorithmic Bias," *Tennessee Law Review* 88 (2021): 649–724; Daniel E. Ho and Alice Xiang, "Affirmative Algorithms: The Legal Grounds for Fairness as Awareness," *University of Chicago Law Review Online* (2020), https://lawreviewblog.uchicago.edu/2020/10/30/aa-ho-xiang; Jason R. Bent, "Is Algorithmic Affirmative Action Legal?" *The Georgetown Law Journal* 108 (2020): 803–853; and Zach Harned and Hanna Wallach, "Stretching Human Laws to Apply to Machines: The Dangers of a 'Colorblind' Computer," *Florida State University Law Review* 47 (617) (2020).

[40] Jack M. Balkin and Reva B. Siegel, "The American Civil Rights Tradition: Anticlassification or Antisubordination?" *University of Miami Law Review* 9 (10) (2003).

[41] Pamela Kirkland, "For Howard Grads, LBJ's 'To Fulfill These Rights' Remarks Are Still Relevant Half a Century Later," *The Washington Post*, June 4, 2015, https://www.washingtonpost.com/news/post-nation/wp/2015/06/04/for-howard-grads-lbjs-to-fulfill-these-rights-remarks-are-still-relevant-half-a-century-later.

[42] Xiang, "Reconciling Legal and Technical Approaches to Algorithmic Bias."

[43] *Regents of the University of California v. Bakke*, 438 U.S. 265 (1978); *Grutter v. Bollinger*, 539 U.S. 306 (2003); and *Gratz v. Bollinger*, 539 U.S. 244 (2003).

[44] Xiang, "Reconciling Legal and Technical Approaches to Algorithmic Bias."

[45] *Students for Fair Admissions v. Harvard*, 600 U.S. 181 (2023).

[46] Ibid., 22–24, 30–32.

[47] Ibid., 27.

[48] Ibid., 39–40.

[49] Xiang, "Reconciling Legal and Technical Approaches to Algorithmic Bias."

[50] Ian F. Haney López, "'A Nation of Minorities': Race, Ethnicity and Reactionary Colorblindness," *Stanford Law Review* 59 (4) (2007): 985–1063, http://www.stanfordlawreview.org/wp-content/uploads/sites/3/2010/04/lopez.pdf.

[51] Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review* 104 (3) (2016): 671–732; and Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, et al., "Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics," paper presented at the Association for the Advancement of Artificial Intelligence/Association for Computing Machinery Conference on AI, Ethics, and Society, Oxford, August 1, 2022.

[52] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," *ProPublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[53] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang, "What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness," *Proceedings of the 2021 Association for Computing Machinery Conference on Fairness, Accountability, and Transparency*, 249–260, https://doi.org/10.1145/3442188.3445888.

[54] George E. P. Box, "Science and Statistics," *Journal of the American Statistical Association* 71 (356) (1976): 791–799.

[55] A. Feder Cooper and Ellen Abrams, "Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research," *Proceedings of the 2021 Association for the Advancement of Artificial Intelligence/Association for Computing Machinery Conference on AI, Ethics, and Society*, 46–64, https://doi.org/10.1145/3461702.3462519.

[56] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* 366 (6464) (2019): 447–453, https://doi.org/10.1126/science.aax2342.

[57] Ho and Xiang, "Affirmative Algorithms."

[58] Anastasia Kozyreva, Stefan M. Herzog, Stephan Lewandosky, et al., "Resolving Content Moderation Dilemmas between Free Speech and Harmful Misinformation," *Proceedings of the National Academy of Sciences* 120 (7) (2023): e2210666120, https://www.pnas.org/doi/abs/10.1073/pnas.2210666120.

[59] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite, "Stable Bias: Analyzing Societal Representations in Diffusion Models," arXiv, March 20, 2023, last revised November 9, 2023, https://arxiv.org/abs/2303.11408.

[60] Xiang, "Being 'Seen' vs. 'Mis-Seen.'"

[61] Salomé Viljoen, "A Relational Theory of Data Governance," *Yale Law Journal* 131 (2) (2021): 573–654, https://www.yalelawjournal.org/pdf/131.2_Viljoen_1n12myx5.pdf.

[62] Antoaneta Roussi, "Resisting the Rise of Facial Recognition," *Nature* 587 (2020): 350–353, https://www.nature.com/articles/d41586-020-03188-2.

[63] See, for example, Kashmir Hill, "Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match," *The New York Times*, December 29, 2020, https://www .nytimes.com /2020/12/29/technology/facial-recognition-misidentify-jail.html; Kashmir Hill, "Wrongfully Accused by an Algorithm," *The New York Times*, June 24, 2020, https://www.nytimes .com/2020/06/24/technology/facial-recognition-arrest.html; and Elisha Anderson, "Controversial Detroit Facial Recognition Got Him Arrested for a Crime He Didn't Commit," *Detroit Free Press*, July 10, 2020, https://www.freep.com/story/news/local /michigan/detroit/2020/07/10/facial-recognition-detroit-michael-oliver-robert -williams/5392166002.