Machine learning (ML) is transforming society as it progresses from research into the real world. There is potential for great benefit, but there are also significant concerns regarding the **fairness** of such uses: disparate performance, unfair job denials, stereotypical search results, and more. So far we have tended to think about fairness in terms of *equality*, which is mathematically convenient to formulate (e.g., as equal accuracy rates across groups). However, I believe we now need to move closer to *equity* and recognize that different people have unequal circumstances warranting different needs. **Fairness as equity accounts for societal context and historical injustices, and is thus far harder to quantify and operationalize** in machine learning models. Yet it is critical that we do so, and in my work, I leverage my technical expertise to engage with these normative questions in ways where they can be better realized in the real world. I do so by: (1) reorienting technical ML research by interrogating the neglected moral concerns behind canonical assumptions; (2) grounding technical work in disciplines outside of computer science like psychology that have a long history of studying inequality and harm; (3) confronting practical issues that arise in the research to reality gap. I thus use my **technical expertise to engage with and unify the work of computer scientists, social scientists, and real world practitioners to develop realistic and impactful ML fairness interventions.**

Reorienting ML research by interrogating canonical assumptions

One of the first things a computer science education covers is the abstraction layer, a powerful tool that hides away implementation details to allow focus on modular components at a time. While this framing has been critical in engineering goals, it has shown itself potentially harmful when applied in the fairness setting [10]. In my work, I interrogate the technical assumptions underlying different abstraction layers in order to inform fairness interventions.

A straightforward layer of abstraction existing in the machine learning pipeline is that of the dataset, which is often taken as a stagnant given. I have worked on complex analyses of datasets including investigating geographic differences in visual representation that could be the result of a tourist photographing a region rather than a local [17, 15], finding simple heuristics such as the primary RGB color in an image to be predictive of perceived gender expression [8], and exploring how image captions describing people of different skin tones differ in racial descriptors across human annotators and AI captioning models [22]. Differences in group representation are not inherently harmful nor even unexpected, but by being precise about the types of dataset difference we find, we can better direct downstream mitigation as well as intervention efforts.

One concern though regarding biases in datasets is that pretrained models trained on them will be biased, with an assumption that this bias will pass on to finetuned models. Given how prevalent transfer learning is, this assumption has sparked great concern. In our work, we first clearly distinguish between two kinds of "bias" that are most prevalent in computer vision: spurious correlations and underrepresentation of groups. We then show that while pretrained models can propagate both kinds of bias into the finetuned model, this bias can be overcome through relatively minor interventions to the finetuning dataset (e.g., recollecting or reweighting 10% of the 128 or 1024 samples), and often with a negligible impact to performance. **Given the ubiquity of foundation models, our critical interrogation of this prevalent assumption better directs attention to the finetuning dataset as the site of intervention** [20].

Another abstraction often used is with the evaluation metric. Metrics like AUC and F1 score are often used without looking to the formula at each use, as is reasonable. Similarly, in fairness, a popular metric is bias amplification [23]. However, in dissecting the mathematical formulation behind the seemingly straightforward measure of bias amplification that was canonically used, I found a number of assumptions: the demographic attribute is binary, each demographic group is uniformly represented in the data, and the task (e.g., objects in an image) and attribute (e.g., gender of the person in an image) are both predicted and entangled by a model. **These violated**

assumptions have serious implications, and we show how results on this metric can produce misleading accounts of bias amplification. To ameliorate these issues, we propose a new mathematical formulation that does not make any of these harmful assumptions [19].

Of course practically speaking, abstraction layers and assumptions are always necessary and cannot be done away with. However, **with my technical expertise and normative understanding, I am able to select the abstraction layers that I believe will have the highest fairness impact.** In an effort to share these problems which require both technical and normative knowledge to a broader community, I have helped to organize popular workshops on Responsible Computer Vision at CVPR and ECCV, two premier computer vision conferences attended by computer scientists, as well as a tutorial on Fairness in Computer Vision at FAccT, the primary machine learning fairness conference which is attended not only by computer scientists, but also by lawyers, policymakers, sociologists, and many more.

Future Directions. Fairness is of course a sociotechnical problem, and computer science has a critical role and responsibility. I plan to **better map out the technically challenging problems that would be impactful for fairness** in order to work on them as well as share them with the community. While prior work took this on at a high-level [1], I aim to do so at a lower, and thus more concrete level. So far, fairness work at conferences like NeurIPS, ICML, and ICLR is technically interesting and mathematically engaging, but the trade-off has at times been a lack of normative grounding because of the assumptions made. These works have been critical in carving a space for fairness problems in ML, but now that we are developing a more complex understanding of these sociotechnical problems, the community's enthusiasm can be better harnessed and redirected towards more substantive fairness work that makes assumptions which are not so severe as to hamper adoption. For example, a large number of works focus on optimization in the binary classification setting where fairness is defined as group difference in some metric from the confusion matrix. However, this setting is not representative of fairness concerns in society. To more empirically understand the proposals for the field that would be most useful and practically taken up, I plan to study the epistemic culture of ML conferences [3], and ICLR's transparent OpenReview process contains a rich database of submissions, reviews, and acceptance results that serve as the perfect site for such an inquiry.

Grounding technical work in disciplines outside of computer science

Computer science overlaps with many other disciplines like statistics, and the subfield of ML fairness especially does (e.g., law, sociology, philosophy). However, interdisciplinary work can frequently involve "epistemic trespassing" where misunderstandings are taken across disciplinary boundaries and can even propagate widely [21]. And yet, to address the deeply sociotechnical problems in ML fairness, we will need to engage in substantive interdisciplinary work.

In one project we confront the problem of ML fairness assuming binary sensitive attributes [18]. On the rare occasion that multiple attributes are considered (e.g., intersectional identities), this is generally done by mathematically extrapolating to accommodate more groups, rather than accounting for the inherently different social dynamics that intersectional oppression brings. For example, math does not tell us which of all possible identities and axes to include in the data—instead, we can navigate these dynamics by combining knowledge on the historical origins of social categories (e.g., how the Asian Pacific Islander category came to be in the USA), and the empirical practicalities of training on groups which are not sufficiently large. Throughout this work, we **investigate the practicalities of incorporating intersectionality in ML, and look to the social sciences to suggest concrete alternatives like this.**

In another project, I collaborated with a psychologist to understand the harm that can come from machine learning outputs [11]. Research often attributes different harms to different people, but doesn't trace the mechanism for how that harm actually comes about. In our work, **we drew from the psychology literature on stereotypes to actually make concrete what**

it means when we say ML classifications can harm someone. We formulate two tangible harms: *pragmatic* whereby an out-group member changes their cognitive beliefs, attitudes, or behavior about the stereotyped group and *experiential* whereby an in-group member experiences subjective emotional harm akin to that from microaggressions. We conduct human studies and come to a complex understanding of harm that will allow for better-directed interventions than currently exist. For example, while women report more experiential harm on mistakes which reinforce stereotypes, in cases of gender presentation men report more experiential harm on mistakes which *violate* stereotypes (e.g., makeup classified on men) indicating discomfort with threats to masculinity. Our results show the need for nuance in harm measurement, and how those relying on automated patterns would miss these complexities in human psychology.

In tandem with conducting interdisciplinary research [7], I also work to build my own skills in non-computer science disciplines by having audited three graduate Philosophy courses and being an active participant in the campus Asian American studies community by taking a course, discussing in reading groups, and attending seminars. In so doing, I familiarize myself in engaging with the literature of other disciplines, and strengthen ties to these communities to work directly with them to produce grounded research. I also host a biweekly AI fairness reading group with participants from departments like psychology, sociology, and public policy.

Future Directions. Beyond disparate incentives and value structures across academic disciplines, one major obstacle in conducting interdisciplinary work can be knowing what has already been written on a topic, especially when it lives in a different disciplinary home. The same word can mean different things (e.g., "bias" as statistical or social bias), and different words will mean similar things (e.g., disparate treatment vs taste-based discrimination, doxastic wronging vs stereotyping). I plan to build a tool that leverages LLMs' ability to ingest massive amounts of content to help find these disciplinary concepts. LLMs notoriously have trouble citing the right, or even real, papers to back up claims, but this task only relies on the LLM to generate large numbers of possibly relevant keywords to serve as inputs for more traditional resources such as Google Scholar. This is not intended as a replacement to interdisciplinary collaborators, but rather a tool to help even find the right collaborators to reach out to.

In thinking about the limitations of this tool, another direction of interdisciplinary inquiry arises, and in fact would benefit from such a tool. As LLMs increasingly become deployed as information retrieval engines, we should consider the strong epistemic consequences of collapsing a multiplicity of results into a single or small set of results. We can find pointers in the long field of epistemology to understand how this might reinforce existing hierarchies in the kind of knowledge that is valued and propagated. I hope to **collaborate with researchers from philosophy and anthropology to study the epistemic effects of LLMs on society.**

I also hope to apply the lens of procedural versus distributive justice to understand the normative implications of technological developments such as GANs to diffusion models and RNNs/CNNs to transformers. In other words, **as the technical mechanisms underlying the *same* task (e.g., text-to-image) change, how we can use procedural notions of justice to understand whether unique fairness considerations arise.**

Confronting practical issues in the research to reality gap

In ML fairness research, there is an urgent need to bridge knowledge production to concrete impact. Academia is also in a unique position where it is less susceptible than industry to market pressures, so there is more space for this critical work that isn't always profitable. In my work, I have mapped out a taxonomy of AI harms [6] then applied them to concrete models [12]. I have also analyzed whether purported principles such as participatory approaches [4] and replacing human studies with LLMs would actually work well in practice [16].

In one line of work, we target a set of deployed real-world decision-making systems that are causing harm to people [14]. We designate a category called *predictive optimization* that is

defined by three key criteria: uses machine learning, predicts the future (e.g., whether someone will pay back their loan), and makes decisions about individuals. This category is broad enough to cover a large range of actually deployed applications, yet tight enough that we can make precise critiques which undermine the systems' ability to function as they purport to. We compile **seven analytical critiques and eight deployed applications, then show how all critiques apply across the board, severely undermining the use of predictive optimization across domains** like predicting life insurance risk, child maltreatment, and school dropout. However, we do not stop short at just pointing out critiques; we also offer a set of proposals forward such as institutional changes to address the root problem, categorical prioritization [5], and partial lotteries.

In another line of work, we seek to understand a more practical barrier to AI fairness: the lack of investment of resources by companies [13]. Despite the large amount of AI fairness research, there has often been a scarcity in adoption. Through 16 semi-structured interviews, I work to **understand the motivations that companies have for expending resources on AI fairness within the constraints of capitalism in order to unveil the levers at our disposal for increasing the investment.** Some themes that emerge include that consumers can have great power in choosing where to spend capital (e.g., using DuckDuckGo to prioritize privacy); journalists exposing the harms of deployed AI systems serve as large motivators to stay away from decisions that could lead to bad PR; and individuals selecting jobs can show they prioritize companies which value ethics, thus supporting those companies that see an "ethics vacuum" and try to set themselves apart. While I remain concurrently hopeful for more radical and structural changes, in the meantime I believe that there are actions we can take in line with the discovered themes to increase corporate buy-in on responsible ML [2].

I also bridge the gap from research to reality by contributing to policy discussions. I was called upon as an **expert witness to testify** on a Hearing for Facial Recognition Technology before the Information, Privacy and Ethics Committee of the **Canadian Parliament**, and was subsequently cited 5 times in the committee's official recommendation to Canadian Parliament on the potential dangers that can arise from facial recognition technology [9].

Future Directions. Building off findings from my interview study that helped illuminate levers in our control to increase responsible AI adoption, there are a number of research directions I see forward that can help translate existing fairness work into real-world impact. One is the strong desire expressed by practitioners to be able to quantify fairness and responsible AI progress in order to convince individuals higher up in the organization to spend resources. Though fairness can never be entirely captured by a number, **by drawing from measurement theory we can measure fairness in a way that makes it more resistant to ethics-washing**, and does not compromise on its normative roots. For example, by grounding each measurement in concrete harms to individuals as I've done in my prior work [11, 12].

Our findings also indicate that well-meaning individuals can enact change at their companies. These well-meaning individuals often come from marginalized identities, so there is much that can be done in the education pipeline that can diversify the space and make it more welcoming. In addition to diversifying the workforce, in our education we can **empower engineers to prioritize issues of fairness**, not necessarily by becoming experts themselves, which I do not believe should be the purpose of ethics integration, **but rather in genuinely appreciating the difficulty and importance of the problem so that meaningful collaboration with the ethics experts can happen.** I expand on this in my teaching statement.

Conclusion: Computer science research in ML fairness is critical, but the disparate communities working on these problems have been increasingly drifting further apart. Through my past and future efforts in grounding this work in sociotechnical realities, I believe we can better unify the divergent approaches for great impact and equity.

References

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G. Robinson. Roles for computing in social change. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020.

[2] Ruha Benjamin. Viral justice: How we grow the world we want. *Princeton University Press*, 2023.

[3] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The values encoded in machine learning research. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[4] Alan Chan*, Chinasa T. Okolo*, Zachary Terner*, and **Angelina Wang***. The limits of global inclusion in AI development. *AAAI 2021 Workshop on Reframing Diversity in AI*, 2021.

[5] Rebecca Johnson and Simone Zhang. What is the bureaucratic counterfactual? categorical versus algorithmic prioritization in u.s. social policy. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[6] Jared Katzman*, **Angelina Wang***, Morgan Scheuerman, Su Lin Blodgett, Kristen Laird, Hanna Wallach, and Solon Barocas. Taxonomizing and measuring representational harms: A look at image tagging. *AAAI Conference on Artificial Intelligence*, 2023.

[7] Arunesh Mathur, **Angelina Wang**, Carsten Schwemmer, Maia Hamin, Brandon M. Stewart, and Arvind Narayanan. Manipulative tactics are the norm in political emails: Evidence from 300k emails from the 2020 us election cycle. *Big Data & Society*, 2023.

[8] Nicole Meister*, Dora Zhao*, **Angelina Wang**, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. Gender artifacts in visual datasets. *International Conference on Computer Vision (ICCV)*, 2023.

[9] Speaker of the House of Commons. Facial recognition technology and the growing power of artificial intelligence. *Report of the Standing Committee on Access to Information, Privacy and Ethics*, 2022.

[10] Andrew D. Selbst, danah boyd, Sorelle Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2019.

[11] **Angelina Wang**, Xuechunzi Bai, Solon Barocas, and Su Lin Blodgett. Measuring machine learning harms from stereotypes requires understanding who is being harmed by which errors in what ways. *ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*, 2023.

[12] **Angelina Wang**, Solon Barocas, Kristen Laird, and Hanna Wallach. Measuring representational harms in image captioning. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[13] **Angelina Wang**, Teresa Datta, and John Dickerson. Motivating corporations to invest in and prioritize responsible ai. *In progress*, 2023.

[14] **Angelina Wang***, Sayash Kapoor*, Solon Barocas, and Arvind Narayanan. Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. *ACM Conference on Fairness, Accountability, and Transparency (FAccT) and Journal of Responsible Computing (JRC)*, 2023.

[15] **Angelina Wang**, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. REVISE: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision (IJCV)*, 2022.

[16] **Angelina Wang**, Jamie Morgenstern, and John Dickerson. Inherent limitations of llms for identity portrayal. *In progress*, 2023.

[17] **Angelina Wang**, Arvind Narayanan, and Olga Russakovsky. REVISE: A tool for measuring and mitigating bias in visual datasets. 2020.

[18] **Angelina Wang**, Vikram V. Ramaswamy, and Olga Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[19] **Angelina Wang** and Olga Russakovsky. Directional bias amplification. *International Conference on Machine Learning (ICML)*, 2021.

[20] **Angelina Wang** and Olga Russakovsky. Overwriting pretrained bias with finetuning data. *International Conference on Computer Vision (ICCV)*, 2023.

[21] Elizabeth Anne Watkins, Michael McKenna, and Jiahao Chen. The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness. *Parity Technologies, Inc., Tech Report*, 2022.

[22] Dora Zhao, **Angelina Wang**, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. *International Conference on Computer Vision (ICCV)*, 2021.

[23] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *Empirical Methods in Natural Language Processing (EMNLP)*, 2017.