April, 2021

AI Principles to Practice

An Introduction to International Activities

Danit Gal The University of Cambridge

Tina M. Park Partnership on AI

Yolanda Lannquist The Future Society

Adriana Bora The Future Society

Niki Iliadis The Future Society

Mia Shah-Dand Women in AI Ethics™

Olga Afanasjeva GoodAI

Arisa Ema The University of Tokyo

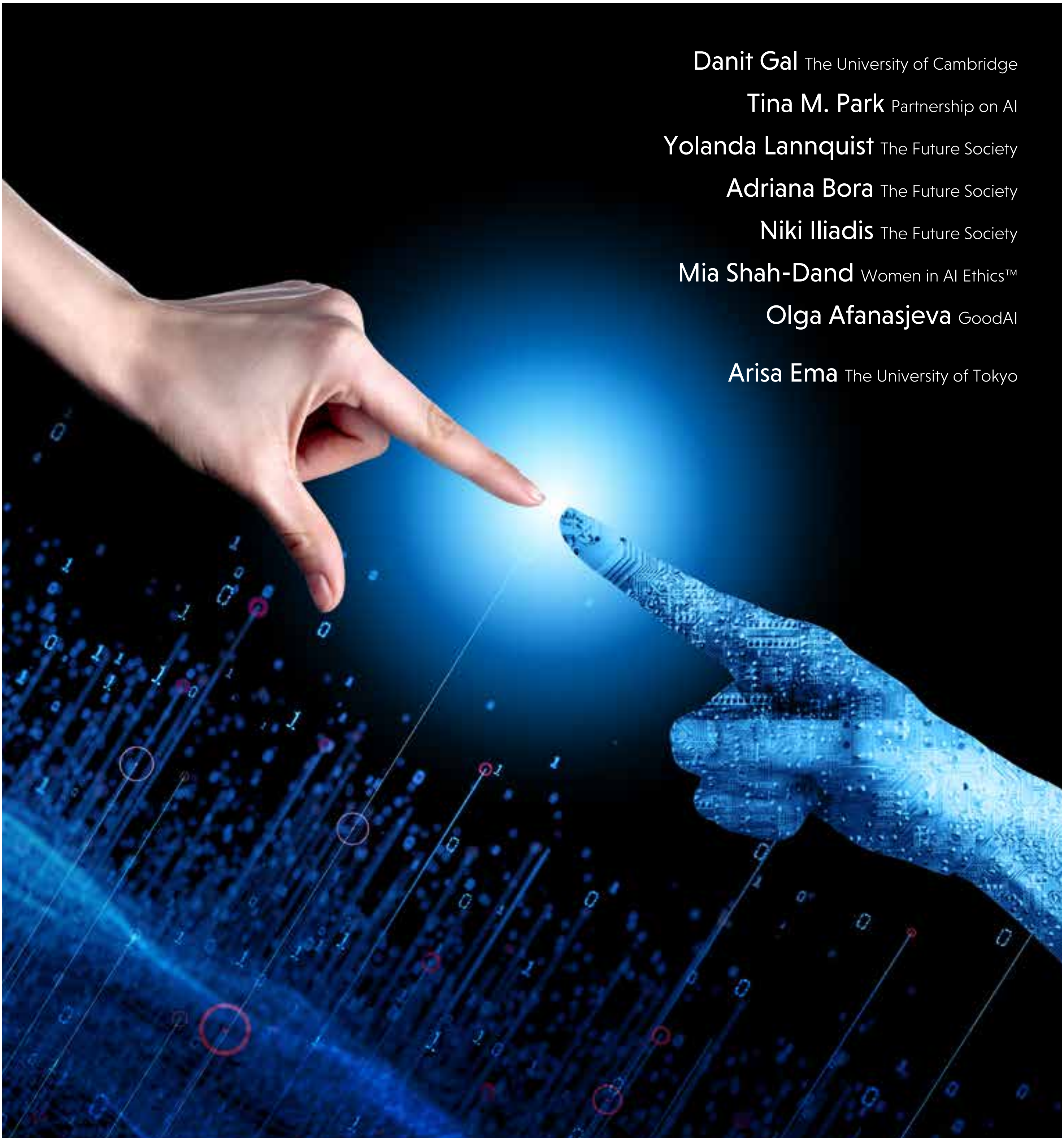
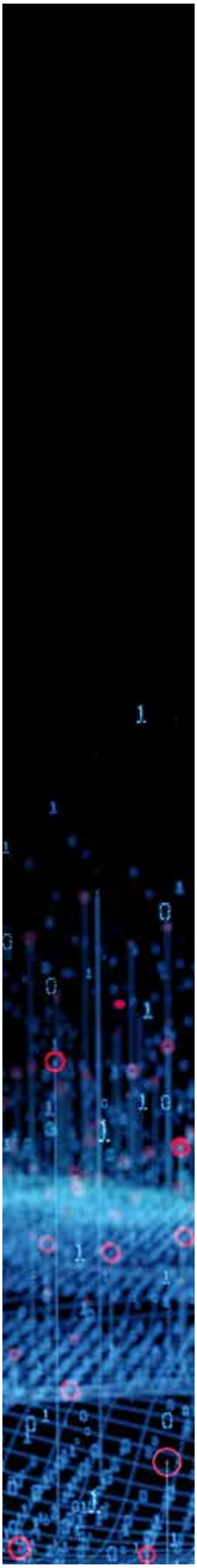


Table of Contents



Paving an Intentional Path Towards Inclusion in AI Development

Tina M. Park Partnership on AI

1 Introduction: Prioritizing Inclusion at the Partnership on AI

Technology holds the possibility of generating both positive and negative effects on the lives of human beings and the world around us. This could not be truer for Artificial Intelligence (AI) and Machine Learning (ML) systems in particular. We are witnessing first hand both the tremendous good enabled by algorithms as we battle the COVID-19 pandemic, like the use of learning-prediction models to identify already existing drugs that could be repurposed to treat COVID-19 [22], as well as their potential for widespread and long-lasting harm. For example, in the United States, Stanford University's Medical Center used an algorithm to determine who would receive the first wave of COVID-19 vaccines, resulting in the exclusion of nearly all 1,300 Resident Physicians working on the frontlines of the pandemic for the hospital [13]. As this global pandemic enters its second year, misunderstandings of what algorithms are, how they work, and how they are created may diminish already weakening trust in public health and vaccine management, as people worry their lives are in the hands of mysterious "black boxes."

The Partnership on AI (PAI), a non-profit organization based in San Francisco, CA, is working towards a future where Artificial Intelligence empowers humanity by contributing to a more just, equitable, and prosperous world. PAI does this by bringing together diverse voices across global sectors, disciplines, and demographics, creating a trusted forum where practitioners and others can share ideas and practices for Responsible AI.

With nearly 100 Partner organizations, including major global technology companies, research centers, and human rights organizations, PAI creates venues to tackle difficult questions about the social impact of AI and ML technologies through both dialogue and data-driven research. In addition to facilitating conversations between experts and leaders in industry, academia, and civil society, PAI conducts research to produce impactful, evidence-based guidance for Partners, and the technology industry more broadly, on how to navigate some of the most pressing concerns related to AI and society.

For example, there has been growing concern about the lack of diversity among technology workers, particularly highly paid engineers and management-level leaders [27]. In addition to reflecting racial and ethnic bias and discrimination in hiring in the technology industry, as well as other barriers to entry in the sciences, the lack of diversity in the AI field is worrisome as it may lead to significant racial and other biases encoded within algorithms [16, 27]. In partnership with DeepMind, PAI launched a diversity, equity, and inclusion (DEI) research study focused on the experiences of women and other minoritized individuals in AI in order to better understand why non-male, non-White employees are leaving the AI sector in disproportionately high numbers and to provide guidance on creating more inclusive environments for those working in AI [7].

Another important area of concern for PAI is the inclusion of diverse voices in the development and deployment of AI and machine learning systems. Our newest research project, Methods for Inclusion, uses a multidisciplinary approach to identify approaches and practices that can be implemented by AI/ML developers and researchers to expand the perspectives and needs considered in the creation of AI/ML technologies.

2 Why Make Artificial Intelligence More Inclusive?

Inclusion is an important tenet of AI/ML development for several reasons. The most obvious benefit of an inclusive approach is the ability to expand who is served by (and who purchases) any given product or service. In other words, there is a business case to be made for inclusion. Even the best products or services are not usable or relevant to everyone; thus, adaptations need to be made to accommodate potential users with different needs. For example, while most Sony Playstation users may find the controller comfortable and easy to use, someone who lacks full mobility and use of both hands is unlikely to play games using that same controller [15]. Adaptations to the controller, or even game play itself like using speech command instead of a physical controller, expands the possible pool of Sony Playstation users to a much wider audience [3].

Thinking about inclusion and exclusion can also serve as a catalyst for problem-solving and the creation of solutions that are better for everyone. An often-repeated example of the benefits of inclusive design in urban planning, for example, is the "curb cut" [4]. The curb cut is that familiar dip in the raised sidewalk that creates a gentle slope to meet the street. The ubiquity of curb cuts in the United States is largely due to the activism of wheelchair users and other people with disabilities. Initially implemented to allow people in wheelchairs the ability to transition smoothly from the sidewalk to the street without assistance, other people found these curb cuts to be extremely useful too. People pushing strollers or heavy carts, travelers with wheeled suitcases, and other able-bodied pedestrians found curb cuts to be very useful additions to their environment. A feature initially designed with a specific audience in mind people in wheelchairs turned out to be an improvement for many others.

More importantly, the inclusion of diverse perspectives, particularly those representing non-white racial-ethnic identities [8, 23], non-male gender identities [5, 10], and experiences of the disand differently-abled [14], is a means to mitigate some of the harm that AI and machine learning systems are shown to cause on already disadvantaged and oppressed communities. In other words, inclusive development, design, and deployment of AI/ML systems may prevent further social harm and help lessen existing social inequalities.

Developing AI/ML systems that are free of social harm is no easy task. For example, Twitter recently came under fire because the ML-based algorithm the company used to crop images on its platform favored white faces over those of Black-identified people [19, 20]. Relatedly, in 2019 several women AI researchers also flagged bias in Twitter's cropping, having identified numerous instances when the faces of women were cropped out of preview thumbnails, focusing instead on their chest [2].

Aware of systemic bias in its own AI/ML systems, Twitter actively attempts to test for gender and racial bias in its algorithms. Yet, despite conducting bias analyses on the ML-based cropping system [1, 12], the company failed to identify this issue until a user came across it, two years after the ML-based cropping system was implemented. As a company, Twitter is also known for intentionally trying to diversify its employee base [6]. While broadening the diversity of perspectives among its engineers is a useful first step in mitigating bias, in this case it was not enough to identify the problem before the image-cropping algorithm was deployed. It required active participation of a broad base of users (and those concerned with bias and discrimination). Although it took a concerted effort of concerned users to finally draw Twitter's attention to the issue, the company did take the feedback seriously and re-examined their algorithms [1]. The responsible creation and deployment of AI/ML systems requires the participation of users, as well as those otherwise impacted by the technology, to design, develop, test, and improve the technology and effectively mitigate any social harms that might result.

For these reasons, PAI has worked since its earliest days to find ways to seek out input from marginalized communities and stakeholders who are not traditionally consulted during the AI/ML development process.

3 Combating the Inclusion Illusion

Technology developers have long been thinking about how to address these barriers to inclusion in their work. Participatory design approaches used in technology development have been around since the 1970s, relying on different stakeholder engagement practices such as interviews, focus groups, user surveys, and system evaluations. Applications from the field of User Experience (UX) research are an important way for companies to understand how someone uses, interacts with, and generally experiences their product or service. Whether beta testing a new product with a set of trusted users or conducting focus groups with potential customers to get a better idea of what users want and need in the latest iteration of a product, this collaboration between developer and user is key to producing and launching a successful AI/ML product (or any product).

At PAI, we first began exploring the idea of working with people outside of the "technical" sphere in partnership with the Tech Policy Lab at the University of Washington, a PAI Partner with extensive expertise in applying value-sensitive design approaches to technology policy [21, 28]. In 2019, PAI worked with the Tech Policy Lab to implement their Diverse Voices methodology within PAI's ABOUT ML project, an initiative focused on establishing documentation practices throughout the AI/ML lifecycle to provide greater transparency to the systems created. The aim was to explicitly solicit views and feedback from communities who are often the least likely to be consulted in the formation of machine learning system documentation practices that nonetheless impact them. The Diverse Voices consultants coordinated three experiential expert panels in Seattle, WA to review and comment on the first draft of the ABOUT ML report.

PAI learned a lot from the careful way the Tech Policy Lab team applied their research methodology towards the aim of greater inclusivity. Specifically, it underscored for us two crucial benefits of incorporating a wider array of perspectives in technology development:

1. It generates important and meaningful insights for tech policy documents, highlighting potential harm or unusability, as well as other uses that were not previously considered.
2. It leverages the expertise of groups that are historically excluded from the development and deployment of technology in mitigating future harm from the use of that technology.

However, it is important to acknowledge that participation is not the same thing as inclusion when it comes to technology development. As demonstrated in other sectors, participation can be used as a disingenuous means to extract labor without proper compensation or credit [25]. Participation may also be used as a way to legitimize the status quo by collecting input without incorporating it into final outcomes and by maintaining boundaries between who is essential versus nonessential in the decision-making process [17, 18]. For example, the Diverse Voices methodology is thoughtful and intentional about respecting the contributions made by experiential experts, or those with a depth of experience and insights gained through life and professional experience, rather than formal education or training. They emphasize the importance of compensating and valuing experiential experts who participated in the panel for both their time and insights. However, the inherent power dynamic between an organization's project leaders and the team implementing a methodology cannot guarantee that a final technology product will reflect the input given by non-technical stakeholders. Additionally, participation itself is not tied to any particular value commitments, other than the belief that more input will result in a better outcome (in this case, product or service). For this reason, metrics of participation often focus on how many people were involved in the development process as a proxy for inclusion. In other words, it is very possible to get lots of "participation" without actually being "inclusive."

Inclusion also requires acknowledging that exclusion exists – that not everyone who should participate in the technology development process is allowed or able to participate to the same degree. Exclusion can occur for many reasons, ranging from a lack of awareness around who else could be included and historic practices rooted in biases and prejudices to institutional policies that explicitly seek to keep certain people out of technology development. Exclusion can also arise, even with a diversity of employees or users, when the contributions of those traditionally with less authority or power are undervalued or otherwise dismissed. This attention to the power dynamics that privilege the needs and opinions of some groups over others is an important distinction between simple participation and full-fledged inclusion. To address these nuances at PAI, we think about inclusion as a form of participation that is specifically oriented towards achieving a sense of integration within a group or institution. Within the framework of "diversity and inclusion," inclusion means creating an environment in which people of many different backgrounds, experiences, and expertise, are involved and empowered to make decisions within the group or organization.

Therefore, achieving a sense of community consultation and empowerment within deployed AI/ML systems requires more than simply soliciting feedback from stakeholders. It requires mapping out pathways through which stakeholders can actually directly engage in the decision-making process, ideally becoming one of many decision-makers and directly influencing the creation and/or deployment of the technology. This means expanding participation beyond whom we normally consider "experts" on AI/ML technology, and identifying how those who hold non-technical knowledge and experiences can become necessary contributors to the development of successful AI/ML systems. It also means cultivating trustworthy relationships between everyone involved so different insights and opinions, including ones that may run counter to existing assumptions, can be freely shared and incorporated into the broader pool of knowledge used to inform the development of a new AI/ML system.

4 Beyond Participation: Methods for Inclusion at PAI

PAI believes that working with communities to develop products and services early and throughout the development process providing multiple touchpoints to assess who is served and how helps AI/ML developers mitigate potential harm or negative impact. A truly inclusive approach can help developers build long-lasting, trusting relationships with the people they want to ultimately serve through their technology.

In order to deepen our understanding of this issue, we created the Methods for Inclusion Fellowship to commit time and resources to research and create materials for those in the AI/ML development community to better capture the nuances of participation and inclusion.

The Methods for Inclusion project is foremost attentive to dynamics of systemic power inequality which have historically resulted in the exclusion and neglect of certain communities and populations from the AI/ML development lifecycle and process. It also extends inclusion beyond the direct participation of individuals representing specific identities or experiences by considering how non-human inputs (e.g., training datasets) serve to include or exclude. This project is multidisciplinary in nature, learning from fields outside of computer science and technology that have grappled with questions of participation and inclusion for many decades. This includes fields like civic governance, education, planning and policy, public health/healthcare, and the social sciences. The project also takes insights and guidance from community organizing, which can cover many topics and disciplines.

Through Methods for Inclusion, we are broadening the aperture in recognition of the other scholarship that exists on the topic of inclusion in various domains. The project builds on lessons we've learned from our valued Partner, the Tech Policy Lab, and the existing work of scholars, practitioners, and most importantly, advocates, who have, for years, tried to open up AI/ML development to people outside of the close circle of engineers and developers.

Throughout 2021, the Methods for Inclusion project will work to:

- identify a range of participatory practices from different fields that could be adapted for use by AI researchers, designers, and developers;
- better understand the challenges of incorporating inclusive methods into AI development, with a specific eye towards the different barriers and incentives facing AI developers, on the one hand, and members of impacted communities on the other; and
- create real-life case study resources that outline attempted participatory methods, the challenges faced by each, and the improvements experienced by companies as a result.

Ultimately, through Methods for Inclusion, we hope to place AI developers and community members around the world who are invested in avoiding potential harm resulting from AI/ML systems into direct conversation with one another.

5 Inclusion as a Global Issue

It is easy to characterize inclusion and exclusion, particularly as it relates to racial bias and discrimination, as solely an issue unique to the United States. However, to believe this would be to deny the histories of migration, immigration, and colonization around the world that have resulted in non-homogeneous societies in every nation. Furthermore, exclusion occurs based on multiple, and often time overlapping, social dimensions such as gender, caste, or social class [9, 11]. Unfortunately, this means that no society is immune to exclusion, bias, and discrimination.

It is also important to recognize that exclusionary values and practices can and do travel. Anti-Black racism is not exclusive to Americans; the belief that Black people are somehow less capable or qualified extends beyond the borders of the U.S., affecting how Black people are treated wherever they may be. Ideas, and the everyday practices and behaviors that emerge from those ideas, circulate globally and embed themselves in organizational contexts far from their site of origin.

For example, Silicon Valley technology companies are currently facing a different kind of diversity issue: the issue of caste. A lawsuit has been brought against U.S-based technology companies for discrimination based on the Indian caste system [26]. Engineers who identify as Dalit, the lowest-ranked caste within India's social hierarchy, allege they experience difficulty getting hired for roles based outside of India because of caste-enforcing practices brought into non-Indian organizations by higher-caste Indian interviewers and hiring managers.

This highlights the importance of considering the presence of exclusion within the AI/ML development process, not only within the local context of one's own city, region, or nation, but throughout the various manifestations of bias and discrimination drawn globally. This is also important because technology itself is mobile. Technology developed in Japan may be used in the U.S., Nairobi, or Brazil. To be an ethical and responsible AI/ML researcher and developer is to recognize that the abuse of technology to deepen social inequality may happen far from where the technology was originally developed. It is important to initiate conversations around how technology not only helps or hinders social inequality from manifesting locally, but also how technology may be used and abused in different social contexts all around the world.

6 Conclusion: Paving Paths Towards Inclusion

Organizations can and should be proactive in their commitment to diversity and inclusion by auditing themselves and their research and development teams to assess the barriers that may stifle contributions made by traditionally excluded communities, such as women and gender non-conforming people, people with disabilities, and racial and ethnic minorities. Who is harmed by exclusion is not fixed and thus identifying those voices may vary from project to project. Working towards inclusion will require careful consideration of the organization itself, the local context (where the development is taking place), as well as the global context (where the technology may be deployed). Intentional pathways should be created so those who are excluded can meaningfully contribute to the design and development of AI technologies. Projects that focus on the social impact of AI technology should be supported and sponsored.

Ultimately, creating many more opportunities for AI/ML developers to learn about and critically examine social discrimination and bias is an important first step in producing responsible and inclusive AI.

Currently, their day-to-day job requirements make it possible for AI/ML developers to create products with wide-reaching global impact that will only compound over time. Thus, it is crucial to equip these technology developers with historical and socio-technical literacy so that they can begin to ask critical questions of the impact of their work and to seek out experts for deeper discussions. Moreover, this responsibility cannot rest solely with the individual employee, but rather must be incorporated through organizational processes (such as oversight, auditing, promotion, etc.). By embedding these practices into the overall organization's functioning, it makes it possible to root out biases and discrimination as a part of day-to-day practice [24]. It also supports individuals to act upon their ethical impulses, whether through whistleblower protections, transparent and responsive decision-making processes, rewards for stopping the release of problematic features, or other mechanisms.

More active and regular conversations about the experiences of women and girls, people who identify as lesbian, gay, bisexual, transgender, or queer (LGBTQ), people with disabilities, or ethnic minorities, led by those with first-hand experience, can only improve the technology that is created. Thoughtful engagement requires being receptive to challenges to your status quo, including accepting that bias, discrimination, and social inequality exist and that everyone, even unintentionally, contribute to the maintenance of these divisions. By being mindful of who is excluded and the extent to which they are excluded from the development, use, and enjoyment of AI technologies, we can actively work towards greater inclusion.

7 Acknowledgments

Many thanks to Kemi Bello and Jingying Yang of Partnership on AI for their editorial and content support.

[1] Parag Agrawal and Dantley Davis. 2020. Transparency around image cropping and changes to come. https://blog.twitter.com/official/en_ us/topics/product/2020/transparency-image-cropping.html

[2] Anima Anandkumar. 2020. I had tweeted in 2019 about @ Twitter cropping #womeninAI headless while cropping men correcting. When I raised this, many men in field accused me of making up a non- existent issue just to gain attention. Sadly #AI #bias is not yet fixed. https://twitter.com/AnimaAnandkumar/status/1307465594304974848

[3] Jason M. Bailey. 2019. Adaptive Video Game Controllers Open Worlds for Gamers with Disabilities. The New York Times (Feb. 2019). https://www.nytimes.com/2019/02/20/business/video-game- controllers-disabilities.html

[4] Angela Glover Blackwell. 2017. The Curb-Cut Effect. Stanford Social Innovation Review (2017). https://ssir.org/articles/entry/the_curb_cut_ effect

[5] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. arXiv:1607.06520 [cs, stat] (July 2016). http://arxiv.org/abs/1607.06520 arXiv: 1607.06520.

[6] Delana Brand. 2020. Inclusion & Diversity Report March 2020. https://blog.twitter.com/en_us/topics/company/2020/Inclusion- and-Diversity-Report-March-2020.html

[7] Jeff Brown and Alice Xiang. 2020. Beyond the Pipeline: Addressing Attrition as a Barrier to Diversity in AI. https://www.partnershiponai. org/diversityinai/

[8] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Conference on Fairness, Accountability, and Transparency 81, 1 (2018), 1–15.

[9] Jane Coaston. 2019. The intersectionality wars. Vox (May 2019). https://www.vox.com/the-highlight/2019/5/20/18542843/intersectionality-conservatism-law-race-gender-discrimination

[10] Sasha Costanza-Chock. 2020. Design Justice: Community-Led Practices to Build the Worlds We Need. MIT Press, Cambridge, MA.

[11] Kimberle Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. Stanford Law Review 43, 6 (July 1991), 1241. https://doi.org/10.2307/1229039

[12] Chaim Gartenberg. 2020. Twitter plans to change how image crop- ping works following concerns over racial bias. The Verge (Oct. 2020). https://www.theverge.com/2020/10/2/21498619/twitter-image- cropping-update-racial-bias-machine-learning

[13] Eileen Guo and Karen Hao. 2020. This is the Stanford vaccine al- gorithm that left out frontline doctors. MIT Technology Review (Dec. 2020). https://www.technologyreview.com/2020/12/21/1015303/ stanford-vaccine-algorithm/

[14] Aimi Hamraie and Kelly Fritsch. 2019. Crip Technoscience Manifesto. Catalyst: Feminism, Theory, Technoscience 5, 1 (April 2019), 1–33. https://doi.org/10.28968/cftt.v5i1.29607

[15] Kat Holmes. 2018. Mismatch: How Inclusion Shapes Design. The MIT Press, Cambridge, MA.

[16] Ayanna Howard and Charles Isbell. 2020. Diversity in AI: The Invisible Men and Women. MIT Sloan Management Review (Sept. 2020). https://sloanreview.mit.edu/article/diversity-in-ai-the- invisible-men-and-women/

[17] MSI Integrity. 2020. Not Fit-for-Purpose: The Grand Experiment of Multi-Stakeholder Initiatives in Corporate Accountability, Human Rights, and Global Governance. Technical Report. MSI Integrity.

[18] Natasha Iskander. 2018. Design Thinking Is Fundamentally Conser- vative and Preserves the Status Quo. Harvard Business Review (Sept. 2018). https://hbr.org/2018/09/design-thinking-is-fundamentally- conservative-and-preserves-the-status-quo

[19] Kim Lyons. 2020. Twitter is looking into why its photo preview appears to favor white faces over Black faces. The Verge (Sept.

2020). https://www.theverge.com/2020/9/20/21447998/ twitter-photo- preview-white-black-faces

[20] Colin (colinmadland) Madland. 2020. A faculty member has been asking how to stop Zoom from removing his head when he uses a virtual background. We suggested the usual plain background, good lighting etc, but it didn't work. I was in a meeting with him today when I realized why it was happening. https://twitter.com/ colinmadland/ status/1307111816250748933

[21] Lassana Magassa, Meg Young, and Batya Friedman. 2017. Diverse Voices: A How-To Guide Facilitating Inclusiveness in Tech Policy. Tech- nical Report. Tech Policy Lab, University of Washington, Seattle, WA.

[22] Sweta Mohanty, Md Harun Al Rashid, Mayank Mridul, Chandana Mohanty, and Swati Swayamsiddha. 2020. Application of Artificial Intelligence in COVID-19 drug repurposing. Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14, 5 (Sept. 2020), 1027–1031. https://doi.org/10.1016/j.dsx.2020.06.068

[23] Safiya Umoja Noble. 2018. Algorithms of Oppression: How Search En- gines Reinforce Racism. New York University Press, New York, NY.

[24] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2020. Where Responsible AI meets Reality: Practi- tioner Perspectives on Enablers for Shifting Organizational Practices. https://arxiv.org/abs/2006.12358

[25] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2020. Participation is not a Design Fix for Machine Learning. In Proceedings of the 37th International Conference on Machine Learning, Vol. 119. PMLR, Vienna, Austria.

[26] Nitasha Tiku. 2020. India's engineers have thrived in Silicon Valley. So has its caste system. The Washington Post (Oct. 2020). https://www.washingtonpost.com/technology/2020/10/27/ indian-caste-bias-silicon-valley/

[27] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating Systems: Gender, Race, and Power in AI. Technical Report. AI Now Institute, New York, NY. https://ainowinstitute.org/ discriminatingsystems.html

[28] Meg Young, Lassana Magassa, and Batya Friedman. 2019. Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents. Ethics and Information Technology 21, 2 (2019), 89–103. https://doi. org/10.1007/s10676-019-09497-z

Tina M. Park

Tina M. Park, Ph.D. Research Fellow at the Partnership on AI (PAI). Tina leads the Methods for Inclusion project at PAI, a project dedicated to developing evidence-based methodolo- gies to incorporate a more diverse range of stakeholders in the design and development of artificial intelligence. More information about can be found at www. partnershiponai.org. Tina earned her Ph.D. in Sociology at Brown University and her Master's in Urban Planning at New York University.

AI Principles to Practice:

An Introduction to International Activities

April, 2021