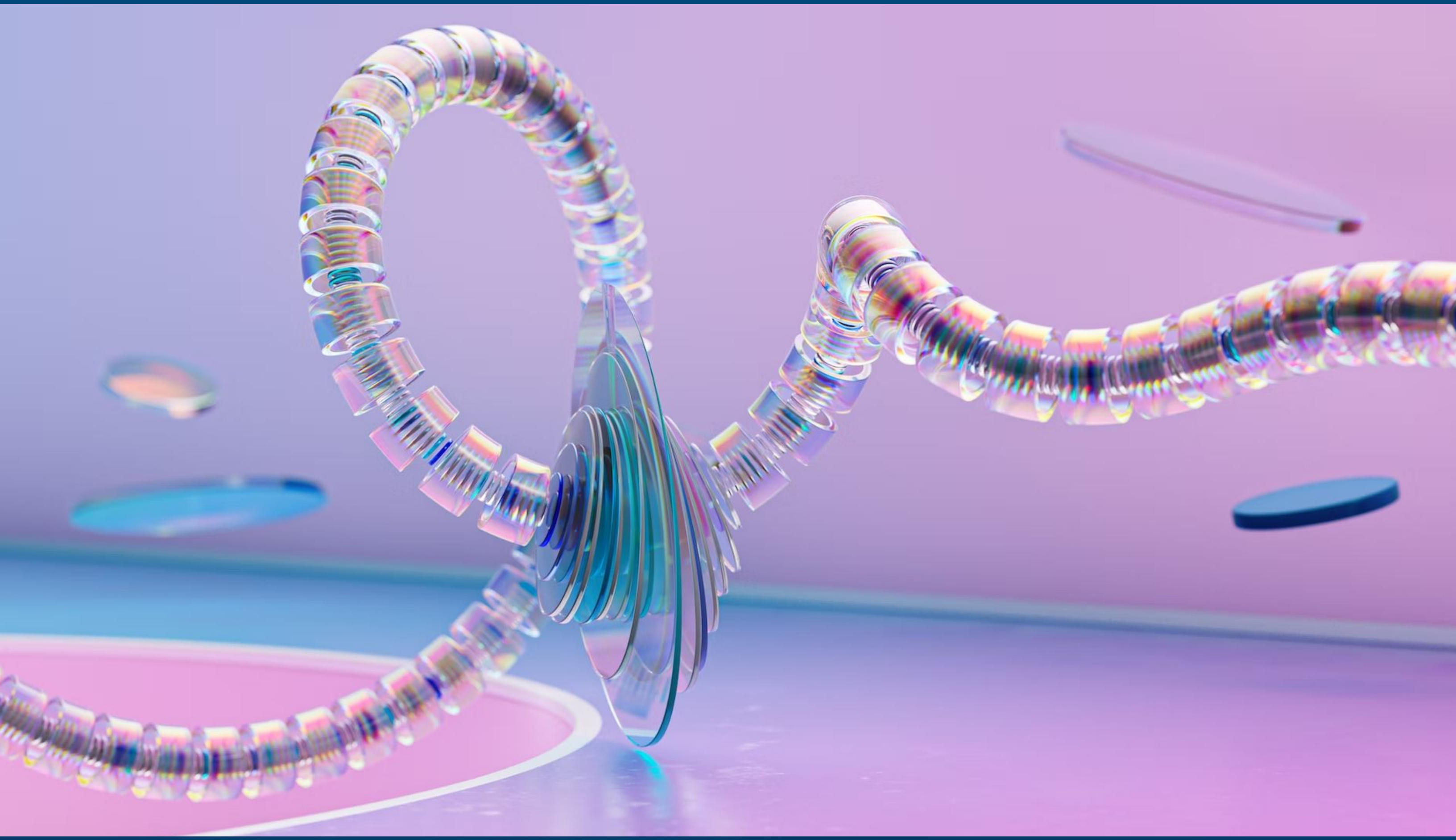
Building and Distributing
Artificial Intelligence for Equitable Outcomes

A Blueprint for Equitable AI

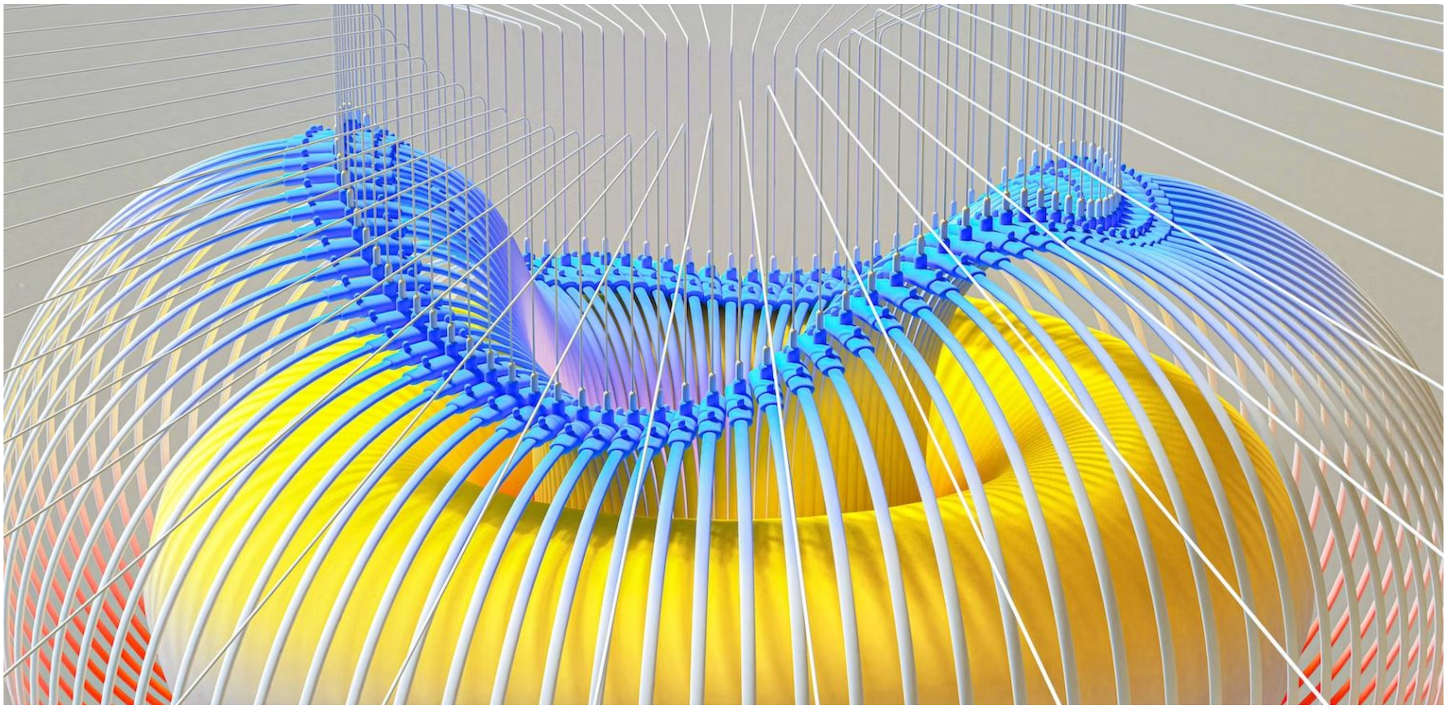


A Report by the
Aspen Institute Science & Society Program

SCIENCE
& SOCIETY
aspen institute

With gratitude for support from

DeepMind



The Aspen Institute is a global nonprofit organization committed to realizing a free, just, and equitable society. Founded in 1949, the Institute drives change through dialogue, leadership, and action to help solve the most important challenges facing the United States and the world. Headquartered in Washington, DC, the Institute has a campus in Aspen, Colorado, and an international network of partners.

In 2019, the Aspen Institute launched the **Science & Society Program** with seed support from the Alfred P. Sloan Foundation and Johnson & Johnson. Science & Society serves as a laboratory to test ideas and approaches that help explain, connect, and maximize the benefits of science for public good. Led by a core staff of trained scientists, the program is an early responder to emerging trends and is on the pulse of critical issues at the intersection of science and society.

THE SCIENCE & SOCIETY PROGRAM WOULD LIKE TO THANK THE FOLLOWING ENTITIES FOR THEIR SUPPORT:

PROJECT SPONSOR

DeepMind

LEAD PROGRAM SPONSOR

Takeda

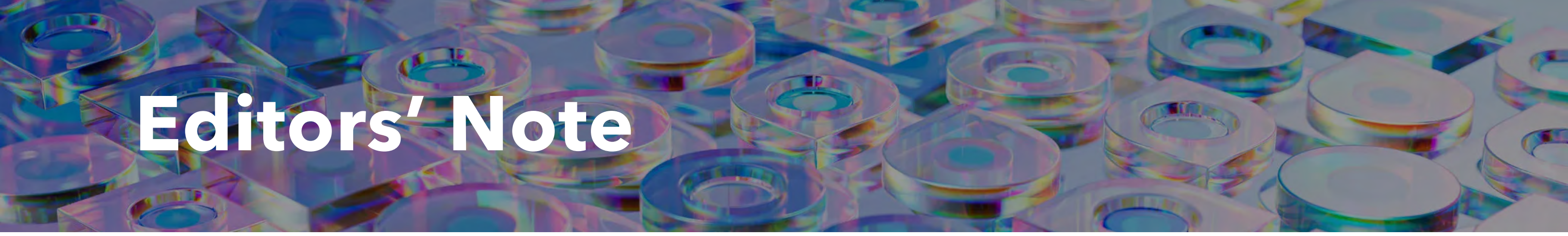
PROGRAM SUPPORT

RITA ALLEN FOUNDATION

Smart Family Foundation of New York
Irma L. and Abram S. Croll Charitable Trust
Nick DeWolf Foundation

CONTACT: Please keep in touch with us at aspeninstitute.org/science, and for questions and comments, please write to Director Aaron F. Mertz at aaron.mertz@aspeninstitute.org.

Table of Contents



Editors' Note

The last decade has seen unprecedented technological advancement, yet grave challenges threaten our global societies. In climate, health and the cost of living, crises are driven by and fuel inequality. Those living in island nations vulnerable to climate change fear the imminent loss of their homes, while the COVID-19 pandemic revealed and exacerbated structural inequities in social, economic, and public health systems. It has never been clearer that we share an urgent responsibility to ensure that scientific and technological advances serve the many, and not just the few.

How should the public and private sectors work together to ensure that historical exclusion does not continue into the present? Pausing technological development and deployment until all concerns are addressed is not feasible—no society has survived without progress. In moving forward, it is critical to ensure that processes and institutions exist to champion and implement efforts toward achieving equitable outcomes.

To that end, we convened two diverse groups of experts in late 2022 to discuss how they might advise building and distributing artificial intelligence for equitable outcomes. In advance of these virtual roundtables, we provided some suggested readings, which are listed at the end of this report, along with a few definitions to start the conversation. This report represents a summary of the discussions.

Many of the ideas explored are not new, nor do the participants offer silver bullets to ongoing challenges. But there is value in exploring them together to build the muscle and future institutions required for civil discourse on the role technology plays in our lives. We hope this report inspires greater curiosity among technologists and the communities they serve, and spurs and shapes the development of markets, norms, and policies toward achieving greater equity.

Sejal Goud – Communications Coordinator, Aspen Institute Science & Society Program

Aaron F. Mertz, PhD – Director, Aspen Institute Science & Society Program

Jylana L. Sheats, PhD, MPH – Associate Director, Aspen Institute Science & Society Program

Dorothy Chou – Head of Public Affairs, DeepMind

With special thanks to **Daniel Porterfield**, President and CEO, *Aspen Institute*, and **Lila Ibrahim**, Chief Operating Officer, *DeepMind*, as well as roundtable participants listed below (alphabetically by last name):

- **Engineer Bainomugisha, PhD, MSc** – Associate Professor of Computer Science, *Makerere University*
- **Chloé Bakalar, PhD** – Chief Ethicist, *Meta*; Assistant Professor of Political Science, *Temple University*

- **Solon Barocas, PhD** – Principal Researcher, *Microsoft Research*; Adjunct Assistant Professor in the Department of Information Science, *Cornell University*
- **Dorothy Chou** – Head of Public Affairs, *DeepMind*
- **Henry Claypool** – Policy Director, *Community Living Policy Center at Brandeis University*; former Executive Vice President, *American Association of People with Disabilities*
- **Lilian Edwards** – Professor of Law, Innovation & Society, *Newcastle Law School*
- **Nicole Foster** – Amazon Web Services Global AI/ML and Canada Public Policy, *Amazon*
- **Rachel Gillum, PhD** – Head of Global Policy, *Salesforce's Office of Ethical & Humane Use of Technology*
- **Tom Lue, JD** – General Counsel & Head of Governance, *DeepMind*
- **Chris Meserole, PhD, STM** – Director of Research – Artificial Intelligence and Emerging Technology Initiative & Fellow – Foreign Policy, Strobe Talbott Center for Security, Strategy, and Technology, *Brookings Institution*
- **Dewey Murdick, PhD** – Director, *Georgetown's Center for Security and Emerging Technology*
- **Sarayu Natarajan, PhD, MPA, LLM** – Founder, *Aapti Institute*
- **Cathy O'Neil, PhD** – CEO, *O'Neil Risk Consulting and Algorithmic Auditing*; Founder, *mathbabe.org*
- **Marie-Therese Png, MEd** – PhD Candidate, *Oxford Internet Institute at the University of Oxford*; Founder, *Implikit*
- **Megan Price, PhD, MS** – Executive Director, *Human Rights Data Analysis Group*
- **Kanjun Qiu, MS** – CEO & Co-founder, *Generally Intelligent*
- **Emily Reid, MS** – CEO, *AI4All*
- **Carol Rose, JD, MS** – Executive Director, *American Civil Liberties Union of Massachusetts*
- **Elissa Strome, PhD, MS** – Executive Director, *Pan-Canadian Artificial Intelligence Strategy at Canadian Institute for Advanced Research*
- **Sandra Wachter, PhD, MSc, MJur** – Professor of Technology and Regulation, *Oxford Internet Institute at the University of Oxford*
- **Theresa Züger, PhD, MA** – Head & Research Group Lead for Public Interest AI, *AI & Society Lab at Alexander von Humboldt Institute for Internet and Society*
- **Anonymous journalist** specializing in technology and society with a focus on China

About Visualising AI

Visualising AI is an initiative by DeepMind that aims to open up conversations around AI. Commissioning a diverse range of artists to create open source imagery, the project seeks to make AI more accessible to the general public. The project explores the roles and responsibilities of the technology. It weighs up concerns and the societal benefits in a highly original collection of works by world-class creators.

The commissioned artworks provide a glimpse into various topics about AI, hoping to encourage further conversation around them. Each artist takes on a theme to transform into unique imagery. The subject matter includes robotics, neuroscience, ethics, safety, methods of machine learning, and more. Experts in these fields are paired with the artists to offer further insight. Crucially, the artists' process is interference free—each artist has creative freedom to visualise the themes however they see fit. From the abstract to the literal, each image is an authentic representation of the artist's take on AI.

To make the collection as readily accessible as possible, DeepMind partnered with Unsplash to distribute the artworks under an open-source licence. Unsplash is an image sharing platform fuelling creativity through a library of millions of pictures. Its users have downloaded over four billion images to date. Anyone from anywhere with internet access and a device can join the Unsplash community for free and enjoy its vast banks of inspiration. Find out more at: https://visualisingai.deepmind.com/.

Featured Artists

Front and back covers: Wes Cockx

Headings: Wes Cockx

Page 2: Khyati Trehan

Page 12: Khyati Trehan

Page 14: Champ Panupong Techawongthawon

Page 16: Vincent Schwenk

Page 18: Tim West

Page 19: Domhnall Malone

Page 20: Nidia Dias

Page 21: Tim West

Page 26: Nidia Dias

Page 28: Domhnall Malone

Page 29: Khyati Trehan

Executive Summary

The two cross-sector convenings revealed several key lessons and themes. Anchoring the conversation were questions of definitions, access, scale, incentives, education, participation, and law.

This report is structured around key questions participants were asked to consider in the context of building and distributing equitable AI, and the topics they led to.

On the topic of defining equitable AI, the discussion highlighted that a fair amount of work remains, including ongoing debates around what technologies count as AI, the need for alignment versus precision, the types of justice being sought, and the potential dangers of a streamlined definition. Experts acknowledged the importance of the specific language used in any definition of equitable AI and who it benefits, while also expressing the need to understand past harms and injustices and to be mindful of local context and history. Participants shared that ideal definitions might:

- Embrace communal diversity, inclusion, and belonging principles at all stages of the process (from design to deployment and beyond);
- Call for transparency and allow communities to know when they are being harmed;
- Keep in mind accessibility and value for all.

The discussion explored the question of **who ultimately makes decisions** around definitions of equitable AI. Questions of attitudes and agency were relevant as participants sought to identify current challenges, strategies, and key players in the push for equitable AI. These included:

- Incentivizing people and organizations who do not see themselves as directly impacted or who benefit from the status quo to care;
- The non-mutually exclusive approaches of incentivizing through morals/ethics, legal compliance, and collective interest;
- Opening up discussions around data sourcing and data access;
- Considering the relationship between the AI and tech industries, market logic, and professional ethics;
- Involving members of marginalized communities in the process, so that their visions for equitable AI are more clearly and tangibly communicated to developers, while ensuring that this work takes place without undue burdens and tokenization.

Emerging challenges can also be the result of processes beyond individual control. For instance, geopolitical factors may complicate dispersed processes of AI development and deployment, or technologies may be applied inequitably outside of their intended use cases. Here, strategies focused on:

- Tapping into and incentivizing existing opportunities for increased control that developers may not be aware of, like adaptive systems for bringing in new stakeholders and opportunities to expand due diligence;
- Returning control where possible to communities, such as by allowing them to label their own data;
- Setting reasonable expectations around the potential of specific technologies and the pace of change;
- Keeping in mind the roles of lobbying and regulation.

Equity is a value, meaning that achieving equitable AI involves **negotiating divergent value systems**. In doing so, participants reflected on the following actions for AI labs and the societies they serve:

- Gain a better understanding of the current state and futures of the value systems at play;
- Contend with the role of scale;
- Recognize the challenges of auditing across divergent value systems;
- Consider how to embrace decentralized systems;
- Support mediation through human forums, while monitoring the possibility of AI mediation down the line.

The conversation also maintained an eye to the future and began to answer questions of **how young people might be prepared for an AI-enabled future**. Important considerations centered around:

- The importance of broader AI education so that existing inequities (particularly on the industry side) are not perpetuated;
- Education being inclusive across socioeconomic status, genders, regions, and knowledge systems;
- Considering how AI will change the future of employment and the skills needed to keep up with this change;
- The need for more examples of equitable AI in context;
- Further research into the impacts of this technology on young people.

In order to effectively implement equitable AI for the future, participants agreed that the conversation needs to be reframed to show how wins and losses are shared across society. A clearer societal understanding of how biases operate, both inside and outside of algorithms, is also a critical step. Often, this does not take place in the straightforward ways it may be conceptualized in the popular discourse, and can be further obscured by the complex workings of algorithms.

As noted during the discussion, AI literacy might be thought of as a prerequisite to equitable AI. Bridging the gap in understanding around AI involves strategies that span the public and private sectors. Examples of such strategies include:

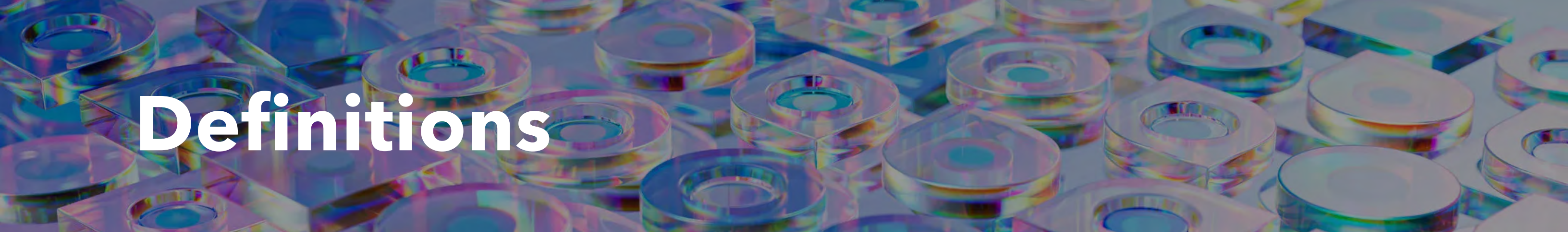
Examples of strategies to bridge the gap in understanding around AI:

- *AI literacy and education through massive open online courses (MOOCs) and universal K-12 curricula;*
- *Engaging the public in direct conversation through town halls and citizens' juries;*
- *AI labs investing in enabling interfaces for users and shaping positive, accessible narratives around AI that encourage public interest;*
- *AI labs allocating resources and internal teams focused on equity;*
- *Forging partnerships with artists and industries beyond AI;*
- *Understanding intermediation and how AI is accessed in various communities;*
- *Simple awareness and transparency.*

The question of equitable AI is one of fairness, and as experts shared, in some respects there is no recourse when fairness is violated. Still, harms can be mitigated by:

- Learning from other high-impacts sectors, such as the relationship between the drug industry and its regulatory bodies;
- Supplementing complaint-based systems, given that AI affects much of the public 'invisibly';
- Fixing grievance redress mechanisms more broadly.

As one participant noted, there will inevitably be mistakes. However, working toward a more equitable future with AI must be a shared responsibility, founded in structural clarity and the instrument of law to cement incentives and good intentions. Despite the transformative nature and potential of AI, several participants shared the view that the solutions brought forward throughout the roundtable can be achieved by incorporating AI into existing democratic mechanisms for technology.



Definitions

Artificial Intelligence

“Intelligence measures an agent’s ability to achieve goals in a wide range of environments.”[1] Artificial intelligence therefore refers to a system with the ability to achieve goals in a number of environments. James Manyika, in the introduction to the AI & Society edition of *Daedalus*, cited the definition “the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings” developed by Ian Goodfellow and colleagues.[2]

[1] Shane Legg, *University of Lugano PhD Thesis*: “Machine Super Intelligence” (2008).

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al., *Neural Information Processing Systems*: “Generative Adversarial Nets” (2014).

Equitable AI

Definitions of equitable AI are commonly laid out against the backdrop of AI risk and responsible AI. However, while responsible AI focuses on the development of AI to “build fairness, interpretability, privacy and security into these systems,” equitable AI generally refers to its deployment.[1] The World Economic Forum, for example, defined equitable AI as focusing on deployment of AI systems in its Blueprint for Inclusion for Artificial Intelligence. It stated there are “growing concerns about bias, data privacy and lack of representation [which means we must ensure] that all affected stakeholders and communities reap the benefits of the technology, rather than any harm.”[2]

[1] *Google*: Responsible AI Practices.

[2] Global Future Council on Artificial Intelligence for Humanity, *World Economic Forum*: “A Blueprint for Equity and Inclusion in Artificial Intelligence” (June 2022).

Building Equitable AI

What details would you add to the definitions of "equitable AI" based on existing literature and your experience in the sector in which you work?

A discussion of suggested definitions for equitable artificial intelligence (AI) revealed a myriad of different interpretations. As one participant noted, it is a definition without historical roots, meaning that it is malleable and might depend on its purpose or application.

Defining Artificial Intelligence & Equity

When shaping a definition of equitable AI, participants expressed the importance of starting with a precise idea of the set of technologies that constitute AI:

- In the United States, this has been a core issue of the White House's Blueprint for an AI Bill of Rights.
- In Europe there have been similar debates about the inclusion of everything from neural networks to Bayesian techniques in the EU AI Act. For some participants, this looked like excluding cases of low-stakes automated decision making and focusing on the highest-impact technologies.
- Meanwhile, others put forth the idea that rather than worrying that definitions of AI technology are being overly inclusive, a definition should take a more holistic approach because those impacted may not know the difference between inequitable AI and biased automated decision making.
- Within the set of technologies defined, how will attention be allocated between traditional, often commonplace use cases like supervised machine learning and the most cutting-edge projects?

Asking questions from an auditing perspective might enable an additional consideration of how AI outputs might be problematic for stakeholders in local contexts, as opposed to attributing inequities to the algorithms themselves—thus helping us to define equity in this context. This audit might take the form of an external risk assessment. For example, are citizens' legal or civil rights being enforced by the AI system? What are specific instances of harm and where are they occurring?

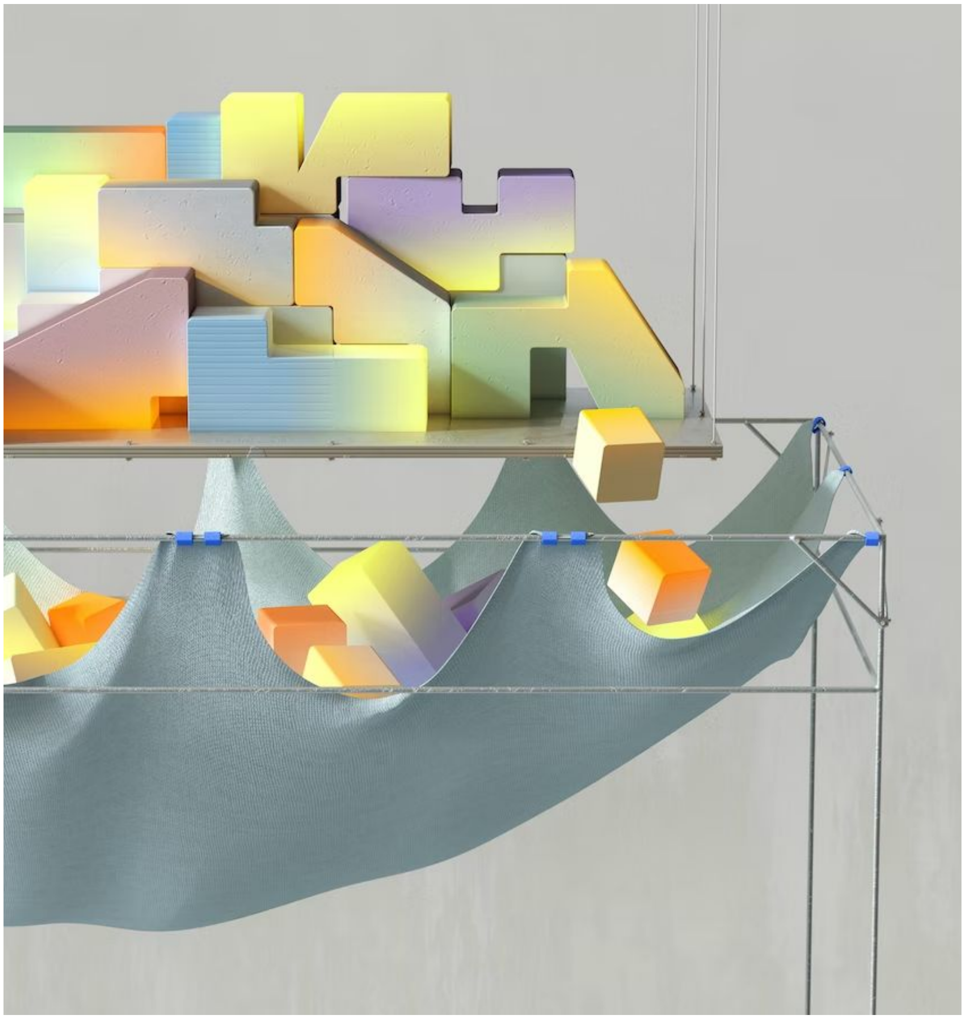
The Importance of Participatory Processes

The discussion arrived at the consensus that one **cannot address equity without an appreciation for diversity, inclusion, and belonging**. Importantly, some stressed that these principles should be considered in terms of collective rights rather than the individualistic perspective that is often the default, reflecting ongoing debates in human rights law.

Here, a focus on diversity must extend to all programming, activities, training, and research. Likewise, inclusion must go beyond the existing goal of ensuring that people are able to **contribute a respected voice** at the table and enters the territory of **access to AI systems themselves**. Efforts should span from inclusive design principles at the earliest stages past the embed deployment of resulting technologies, so that equitable AI considers societal impact rather than the product life cycle alone.

In **prioritizing access and participation**, it becomes possible to envision a world where the outputs of AI are of value to everyone, including those who use different languages and those who might not own high-end smartphones or computer infrastructure. Hence, participants emphasized **AI literacy as a prerequisite** for equitable AI and the importance of thinking about deployment infrastructure.

With regards to literacy, an analogy to AI was made with the Gutenberg press and how futile it would have been if people could not read—even if it were used to produce books in multiple languages. To ensure that all stakeholders are brought into this conversation on a level setting, it is critical to **build capacity for marginalized communities** to better understand the current state of the technology and the path toward a safer future. Without these key voices, even the most well-intentioned projects can become inequitable.



Justice: Incorporating Local Context and History

Referencing the pre-readings, particularly Iason Gabriel's article "Toward a Theory of Justice for Artificial Intelligence" (2022), attendees noted that it is important to ask what justice means in this context. Instead of thinking about harm and justice as singular concepts, the conversation can be further specified through approaches such as distributive, compensatory, transformational, restorative, procedural, and structural justice. In following these avenues, there is much to be recognized and learned from those **already organizing change**, including indigenous communities.

At some level, refining a definition of equitable AI also entails understanding current inequities. For this reason, some stressed that any definition must explicitly allow individuals and communities to **know that an AI tool is being used in relation to them**. Without this knowledge, those outside the industry are powerless to interrogate the system and react to the effects they are experiencing, which may be significant but invisible, such as the denial of a loan as the result of an AI process. While equity entails listening to all stakeholders, **specific attention** must be paid to stakeholders for whom AI is failing rather than unduly focusing on companies who are already making sure that systems benefit them.

Drawing from multiple participants' past experience working with autonomous vehicles, three critical questions were identified:

1. *How safe is safe enough / how fair is fair enough / how equitable is equitable enough?*
2. *How is this measured?*
3. *Who gets to decide?*

The particular intricacies of measuring fairness and equity were also highlighted relative to physical safety, where the single metric of a vehicle being involved in a collision may suffice to identify an unsafe product.

Since AI is deeply **embedded in a society that is already unjust**, that makes establishing a definition even more challenging. Moreover, some participants highlighted that differences by country—including those that might be assumed to be similar such as the United States and Canada—mean that a **global concept of equity** might be needed for the definition of equitable AI to apply across the board. However, whether global concepts like fairness can or even should exist was challenged as the discussion continued.

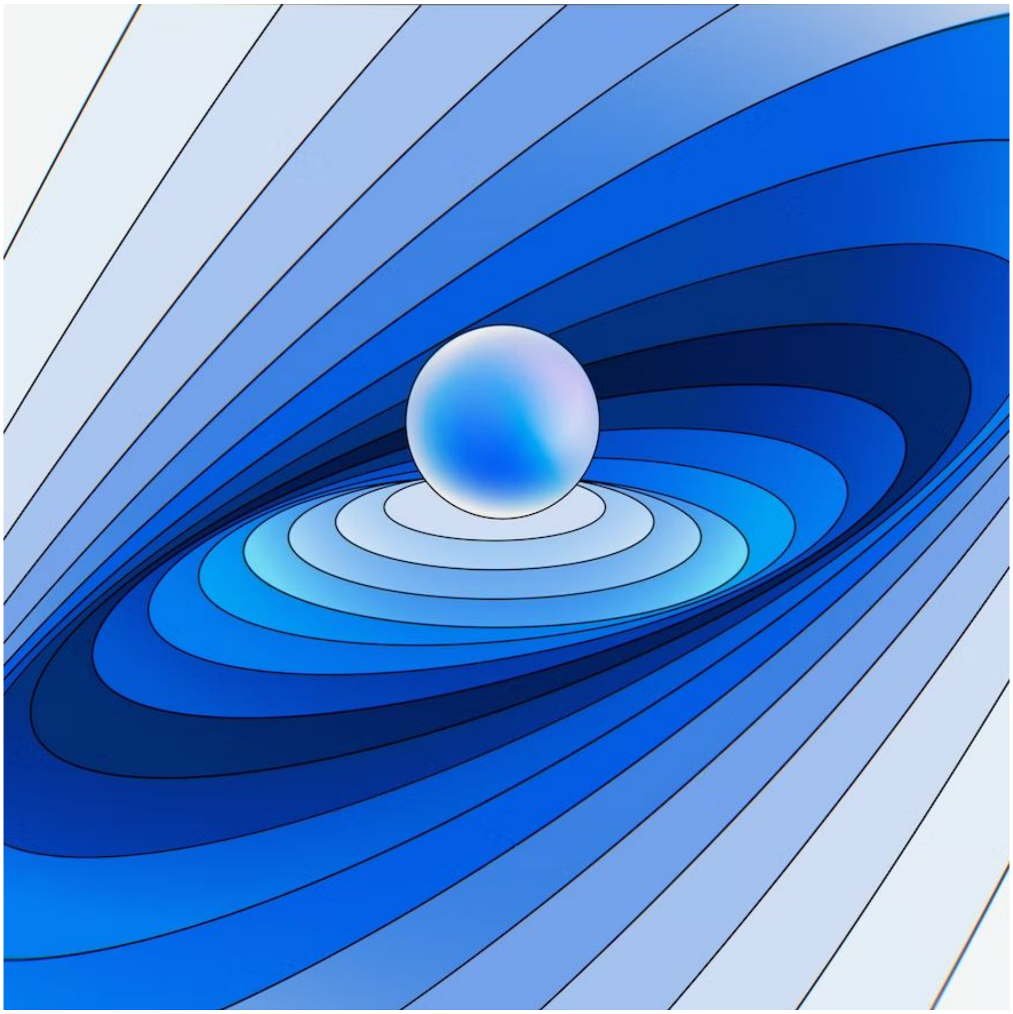
Working Toward Mutually Beneficial Outcomes

Instead of a **perfect definition** of equitable AI, some participants expressed that society is best served by alignment on **common, shared definitions** across teams building the technology and externally in wider society.

At the same time, other participants **challenged** the idea of a **streamlined definition** of any kind, suggesting that operating from a binary standard for fairness caters to private sector actors who seek to automate and scale the definition. Instead, a more desirable definition includes an endorsement that it is not binary and has the **capacity for constant change** and to recognize different histories, jurisdictions, cultures, and legal perspectives. The definition should possess a similar agility to what is demanded of society's legal systems.

Without drawing too firm of a line, some suggested it might be useful to **distinguish between the development and design phase, the deployment phase**—which includes infrastructure and scaling—and **downstream impacts in communities**. As one expert noted, it is possible to achieve equity in these first two stages yet be left with inequitable outcomes. Per another contributor's suggestion, perhaps AI equity might be better positioned as being aligned with Turing's approach, which puts forward a test rather than an actual definition in its practice.

Moving beyond these phases, there is complexity introduced by AI's status as the first type of technology to have its own goals. Currently, only the designer of the algorithm has the ability to determine its goals. While the end users can make contributions, they lack the power to change these goals. In contrast, several participants argued that equitable AI systems should have goals that are **transparent to and controllable by** the end users, at least through bounded delegation.



The Language of Equitable AI

The domains from which the vocabulary of definitions originates also matter as a reflection of societal values. Different outcomes for equity result from keeping discussions of equitable AI **grounded in the legal language of rights** as opposed to the clinicized language of companies and regulatory bodies.

Even when working from a common language, there are bound to be **differences by legal system or jurisdiction**. As one legal researcher noted during the roundtable, the legislation around equity in the United States and Europe often uses the exact same words to mean two completely different things. This can be particularly problematic when companies use bias tests developed using one country's definition of fairness to assure legal compliance in another country, thereby **unknowingly putting their algorithm(s) in conflict with the law**.

To reach consensus between private companies and the public good, it is also important to be cognizant of competing incentives in definitions and ensure the interests of marginalized groups are brought to the fore over definitions that solely speak to profit incentives.

What are current challenges and key strategies for more effectively creating equitable AI? Who are the key players who must work together?

Referencing the opening question of the roundtable, the conversation around challenges quickly pointed to the obstacles presented by lack of alignment on the definition of equitable AI and AI itself. The discussion then identified several additional challenges, involving both public and private sector players.

Inspiring and Incentivizing Collaboration

Moving beyond definitions, participants identified the issue of getting those who are **not directly affected by the inequities of AI** to care. The question of incentives can be particularly challenging in the context of start-up firms who are seeking to stretch their funding. In a competitive venture capital market, this often means that smaller companies cut corners around ethics and safety. At the same time, it is also applicable to those working in big tech.

The way technology currently operates is optimized for populations in power rather than everyone affected by it. In the absence of force, some actors are **unlikely to make meaningful progress toward equity** because they prioritize profits. They may also be quick to lobby against broad enforcement because it is costly to simultaneously achieve accuracy, efficiency, and

equity. In effect, participants broadly agreed that one stakeholder's success may currently be another stakeholder's inequity.

Three strategies for reconciling these conflicting goals emerged from this discussion:

1. *First, incentivizing people from moral and ethical perspectives—that is, telling them why equitable AI is something they ought to stand up for;*
2. *Second, consulting the law and demanding legal compliance;*
3. *Third, making it clear in both the public and private sectors that there is a collective interest in inclusive systems rather than continuing to operate from a point of view that yields suboptimal outcomes for everyone.*

The chosen strategies (which are not mutually exclusive) must prompt those who do not need to invest in this issue to do so. Otherwise, the conversation ends before decision-makers are able to act. Moving the needle also requires recognizing the political nature of these efforts and **incentivizing equitable AI as a political priority**.

Data Sources and Data Access

The question of who has access to the data used to build AI is a central debate across sectors. As participants shared, more open access to better data allows new actors to enter the field.

Currently, many emerging, data-intensive models are developed using **scraped data**. As one participant noted, this is problematic because the resulting models are trained without community-driven labels and often do not represent the global community due to the affluence and geographic concentration of their respective data sources. By building and training AI using American or European infrastructure exclusively, a norm is established where AI infrastructure is predefined to only be available in certain contexts.

Instead, efforts to **crowdsource multilingual data** through international data trusts and data cooperatives represent one path towards more equitable AI. Labs with a global reach might also consider strategies for sourcing data that are less dependent on outsourcing alone.

Looking ahead, greater investments will be required to understand the processes of deep neural networks and how they relate to data in ways that impact safety and equity.

Incorporating Market Logic

From the perspective of one participant with a background in public interest AI, it might be argued that the industry does not currently follow market logic as one would expect. Alternatively, a different perspective suggested that there are drawbacks to aligning the creation of equitable AI with money. Often, the conversation overlooks that people are willing to pay more in exchange for safety.

From the observations of one scholar, even when individuals are committed to equitable AI, there is little change in the **ethical character of companies** because once they enter the work-force, these well-intentioned voices come into contact with dominant corporate business models and can be driven by a desire to fit in with their peers. Alternatively, a participant from the industry side suggested that there is cause for optimism because large public companies are driven to care about their reputations and are in a position to lead or model the way for a shift across the industry.

As many agreed, proposals for a **professional ethics system** similar to what exists for doctors and lawyers are an overly simplistic solution that fail to account for the different structure of reputational capital in the tech space. Particularly since engineers work in groups, individual accountability can be difficult to enforce. Moreover, participants expressed that much of the industry is driven by performance improvements on benchmarks rather than the pursuit of understanding that other scientific fields might prioritize.

Launching AI products in a limited environment and continuing to audit represents one strategy for working around fears of liability that can limit intensive auditing. Additionally, plural governance structures and the creation of dedicated and diverse teams focused specifically on equity within a general culture of equity at AI labs is important. Even still, technologists expressed during the discussion that it can be difficult for a small group to gauge the equity of the AI they are building and deploying. Here, **red teaming and maintaining an open repository of flaws** allows for critical knowledge sharing. However, the success of these strategies hinges on previously discussed incentives.

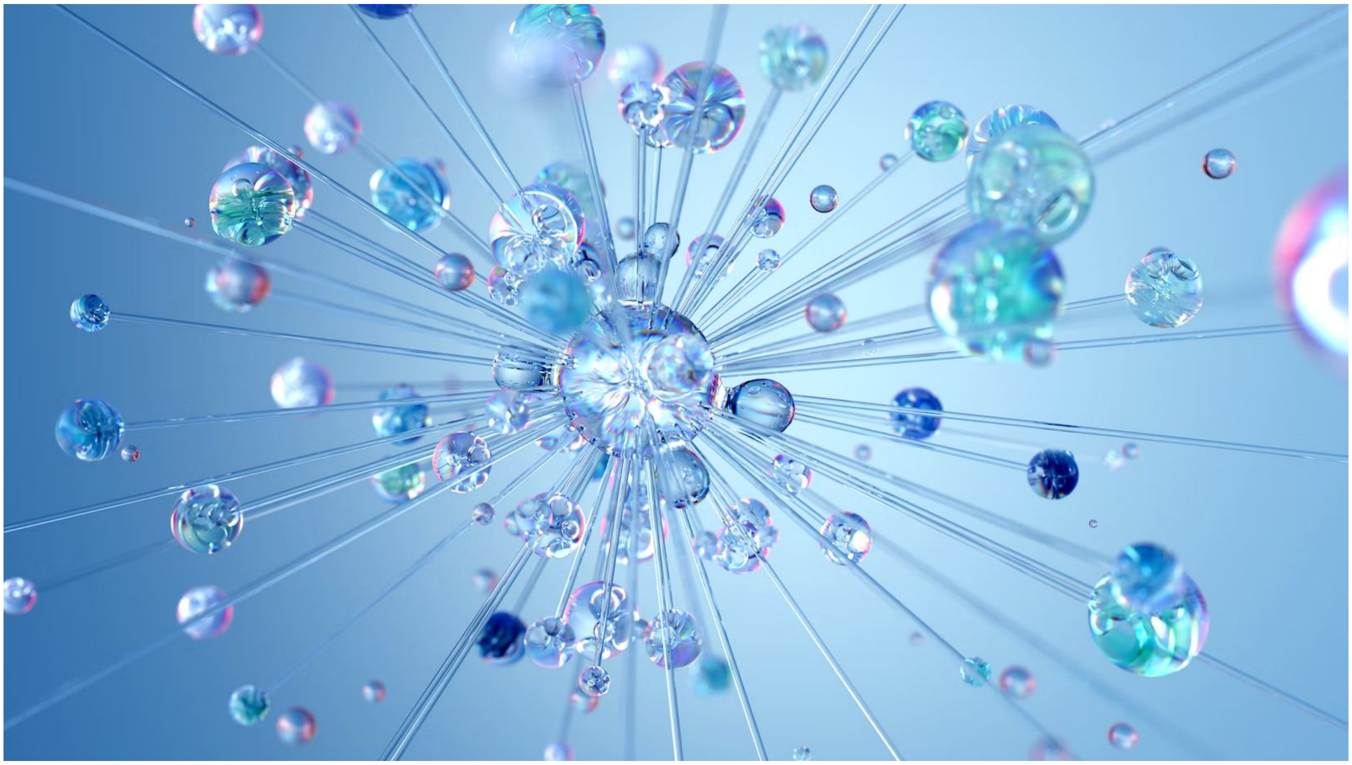
Trust, Agency, and Participation

More specifically, without access, education, and forums for conversation, communities themselves may **lack a clear idea or articulation** of their goals with respect to equitable AI, making it difficult for those trying to design systems that meet those needs. As one expert noted based on their field experience, there is a distance—both physical and in terms of language—between organizations that build AI and convene discussions around it versus the populations who are most affected.

In addition to building capacity for community members to understand AI tools and diversifying internal teams by working with members of affected communities, it is important to **push for deeper collaborative engagement** while being mindful of tokenization. Speakers expressed concern about the ability to find advocacy groups that best represent stakeholders because they are often poorly funded. This includes the need to support voices from civil society that lack a technical background but are **knowledgeable and connected** to vulnerable groups. As one example, including perspectives from the disability community means an opportunity to work with a diverse group while opening up questions about civil protections from the state. Experts also stressed that society must reckon with the **paradox of participation**, so that efforts extend beyond numeric representation and take steps to redistribute power. As was noted during the roundtable, this includes making sure that activists and those who rock the boat feel able to contribute in spaces where the culture is to maintain the status quo.

Participants emphasized that even without technical know-how, **community members can be involved** in the development process. For instance, they may be allowed to serve as data points, share their knowledge of negative use cases as part of red teams, and provide their ideas on how to change applications. Similarly, students at the intersection of marginalized identities

must be involved, as they are often the most affected by AI. However, much like work done to improve diverse representation in other sectors of society, working toward **equity must be a shared responsibility rather than a burden that falls on the shoulders of these young people**. All shareholders in the conversation should support the ascension of marginalized youth into leadership positions in AI in order to facilitate equity.



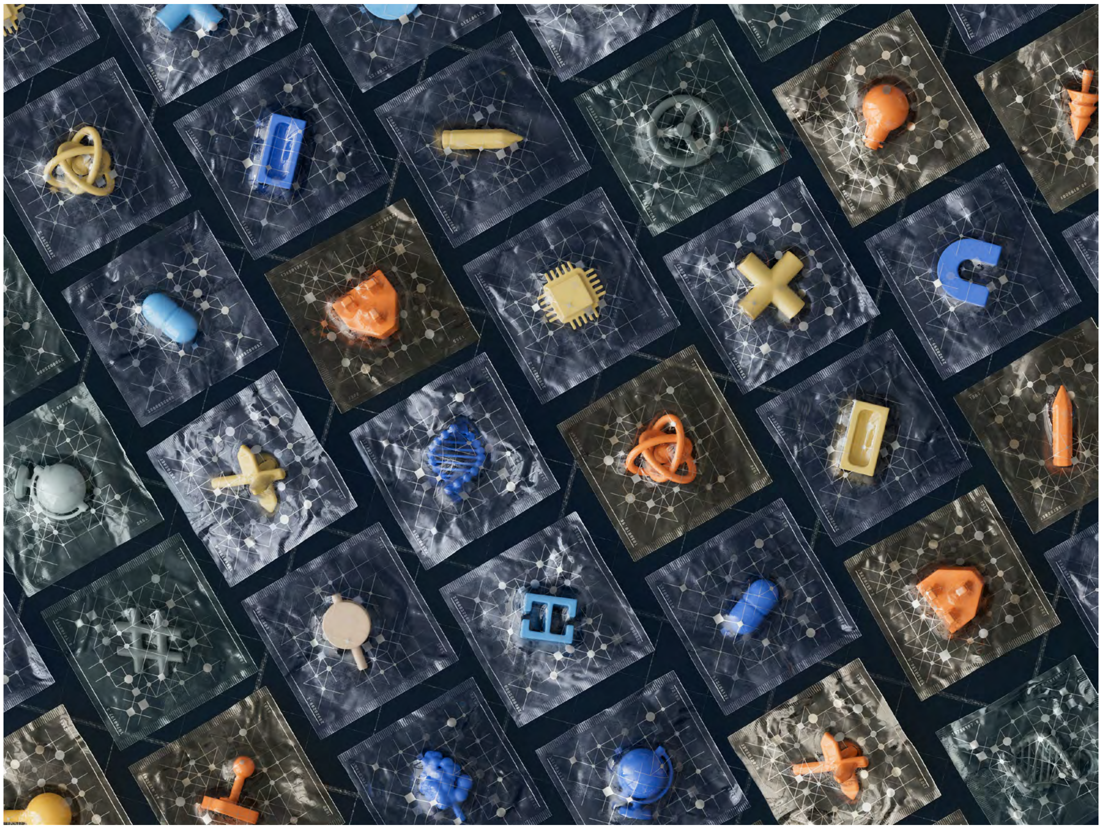
How do you manage external developments that have impacts on equity but are outside of your control?

One participant flagged that the dispersed nature of technological development might make **AI equity susceptible to impacts brought about by geopolitical factors** at all scales, ranging from the local to international level. These factors highlight a lack of control, both for nation-states and individuals.

At the same time, strategies that involve both (i) empowering developers with actionable steps and (ii) communicating how these measures are aligned with their best interests—keeping in mind that this may not be true in the case of a market failure where further policy is needed—represent a **path toward equity by leveraging existing control** that technologists are perhaps not aware of.

Capitalizing on What Is Possible

Participants from multiple sectors voiced that accounting for external developments that impact AI requires **reevaluating how labs conduct due diligence**. There is room to be more thorough and to incorporate additional context that AI designers may not be considering. This includes designing adaptive systems, so that there is an established procedure regarding newly-identified stakeholders and blindspots. Moreover, it is **not sufficient for due diligence to be isolated to the front end of development and to the intended use case** for the AI system. In cases where technologies are being applied in inequitable ways that were not originally antici-



pated or intended, it remains the responsibility of AI labs to **respond, learn, and seek** means of mitigating these harms in the future.

Another recommendation involves focusing on establishing minimal thresholds for datasets to optimize the number of people who will benefit from AI. It is difficult (if not impossible) to articulate what constitutes a ‘good’ data set, and the challenge of how best to collect and license data from disenfranchised communities raise further ethical questions. This is particularly relevant in light of popular data-driven models where AI labs may not be able to fully account for biased representation in data and concerns about data equity. By enabling different groups to label and have agency over their data, companies can move beyond recognition of harm and simple calls for transparency in order to begin shifting actual working dynamics. As circumstances change, there should also be an expectation that there will be a high variance in the data sets used.

Establishing Expectations

The conversation also highlighted the need to **set reasonable expectations** with regards to the ability of individual AI products and organizational restructuring to rectify all issues of equity in the face of historical patterns of marginalization. Without **enforceable laws and regulation**, there is no level playing field. Some posited that if well-intentioned actors in the AI space are also those on the front lines against legislation that supports equity because they fear its economic impacts, it is difficult to realize change.

How can/should AI labs negotiate divergent value systems when it comes to what equity is?

As a starting point, one expert emphasized that more work to understand multiple perspectives must take place through empirical and qualitative research, as well as deep philosophical thinking to build on existing work in value alignment. It is crucial to have a clear sense of what societal values are, where they are and should be going, and what value systems are diverging before making a meaningful attempt at answering the question of how to approach negotiations, while also recognizing that there is no perfect or one-size-fits-all way to negotiate divergence.

Challenges of Serving Society at Scale

When considering divergence, participants pointed out that for many use cases, scale can inherently derail equity, which is **a community-specific and culturally-specific concept**. In theory, the smallest AI systems are poised to have the greatest potential for equity because they are designed in a bespoke fashion. By contrast, the main players in the AI space are seeking and producing systems that cut across billions of people. Several participants agreed that equity is not possible at the current scale of development & deployment, which **forces the selection of one value system over others and universalizes it**. One contributor highlighted the specific example of auditing at scale in relation to the availability of AI skin lightening features on social media apps, which may be considered acceptable in some circles but problematic in others. Perhaps, participants discussed, it is better to focus on asking how to **enable more distributed small-scale development**—a different status quo for AI development.

Relatedly, this question of negotiating value systems takes on new meaning if the assumption that AI is and must be a single thing is challenged. Currently, the centralized process by which many systems are built means that society is largely limited to expressing a single set of laws through AI. Though there are movements away from centralization, a faster shift ought to be prioritized. By embracing decentralized systems, e.g., the fact that AI is not a monolithic concept, society will gain the capacity to express an array of different values by enabling the end user to modify or train the system to serve their own needs and desires rather than those of the company.

As a future and longer-term approach, AI systems themselves might serve as a cost-effective, clonable, and scalable alternative to some labor-intensive negotiations and processes. However, depending on how these systems are built and distributed, there is a possibility that this approach may perpetuate the very issues of scale and inequity discussed previously.

In addition to tackling the issue directly through technology, **ethics advisory councils** provide a forum for negotiating value systems, particularly in corporate settings. Importantly, these groups must meet regularly for conversation, engage internal and external stakeholders, and be treated as a starting point for continual iteration rather than as an end in themselves. As one participant added, the work of these councils should be used to develop transparent policy procedures in advance of crises resulting from differences in value systems, so that AI labs have an intentional process that does not come across as ad hoc brand protection.

Distributing Equitable AI

How do we prepare young people for an AI-enabled future?

One researcher remarked that they did not see a future enabled by AI but rather saw a future in which AI's capabilities could enable young people to create futures they aspire to.

Broadening Education about AI

Enabling young people to thrive in an AI-enabled future is a question of **exposing, equipping, and inspiring as many people as possible to pursue paths in AI**, so that it is not limited to a few privileged individuals. If a more diverse group of young people is not involved, imbalances in AI systems that are ultimately reflected in greater societal imbalances will continue.

Across backgrounds and international perspectives, a common thread running through approaches to preparing young people for AI is **ethical tech education**. Similar to the goals of AI education among adults, **explainability, transparency, and awareness** of existing technologies must go beyond the descriptions of algorithms that some companies currently offer. This education must support youth from all socioeconomic backgrounds, as well as pay attention to those pursuing advanced degrees in computer science who will go on to be developers, including youth whose curricula should highlight the value of both technological ways of thinking as well as indigenous knowledge systems. As an additional strategy, one participant suggested leveraging influencers, many of whom play a critical role in shaping how young people make decisions.

Employing the Future

Discussion of the future cannot omit the topic of employment, where shifts in the labor market as AI develops will bring new opportunities and challenges. Here, participants supported a focus on human capital development through **basic digital literacy, skilling, and training** that later reaches the specifics of the AI system. For this to be successful, sufficient examples of equitable AI in context are required. There is also a need for investment in enabling infrastructure. Data and support systems ought to be in place so that, with the guidance of the education system or other organizations, students have the opportunity to go into the community and build examples of equitable AI systems.

Opportunities for Further Research

As this future nears, further research on the impacts of around-the-clock access to rapidly advancing technology on childhood development is called for. With this knowledge in hand, current players in the technology space can be better equipped to build a future where AI delivers positive social benefits rather than harm to younger generations.

What needs to happen societally in order to effectively implement equitable AI?

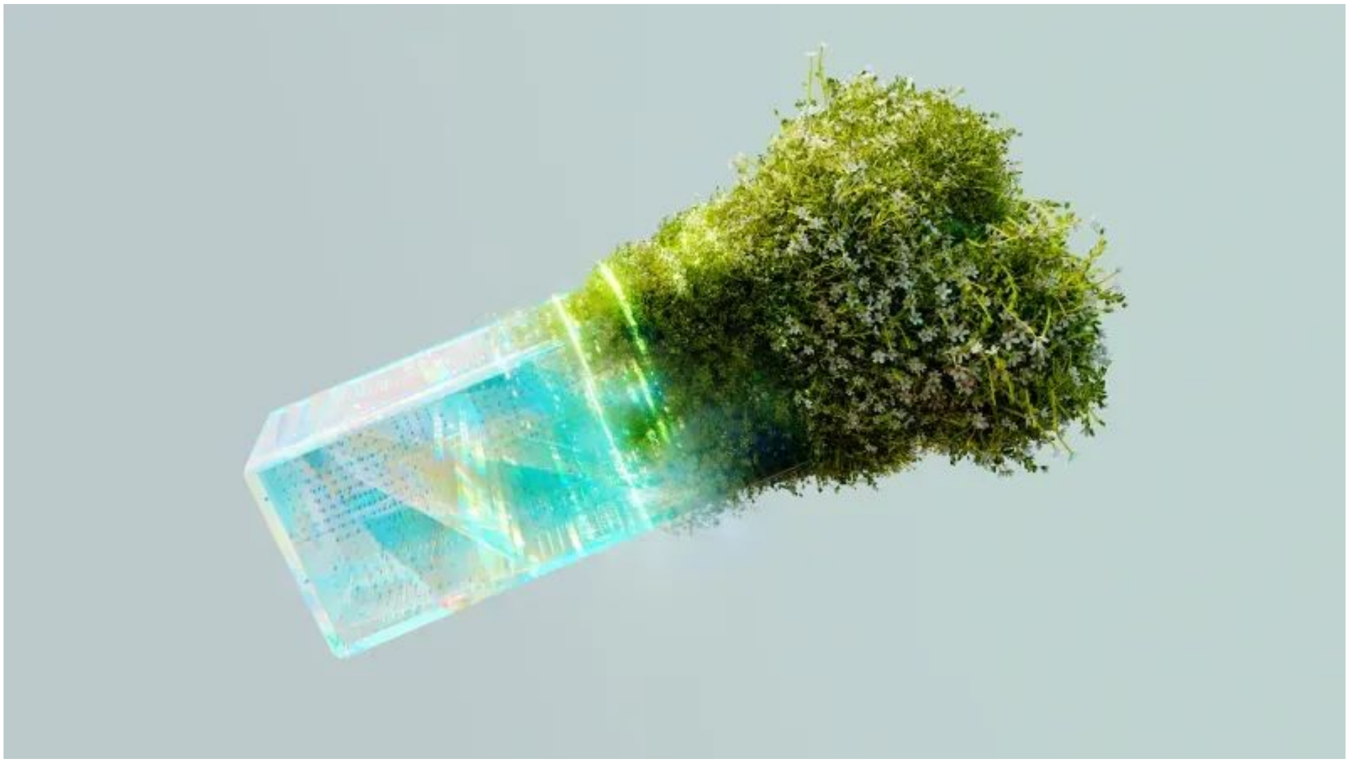
At this stage in the discussion, participants agreed that effectively implementing equitable AI is a matter of generating **broad recognition of the consequences of inequity and understanding of the problem at hand** among the general public.

Gaps Analysis

Parallels can be drawn to the struggle to get people to care about environmental issues, where it unfortunately often takes a severe or deeply personal event for people who see themselves as disinterested parties to join in the conversation and imagine their role in the solution. By framing **equitable AI as critical to decision making in areas of paramount importance such as national security, healthcare, and policing**, society is better able to conceptualize the importance of shifting away from the status quo.

Recognizing the Harms Resulting From Inherent Bias

Achieving equitable AI requires debiasing not only AI systems, but also **addressing the biases present in the humans** who come into contact with these and other systems at every phase. As one expert noted, even in instances where society recognizes systemic discrimination such as racism, sexism, and ableism, these biases are often understood in abstract forms that fail to capture the multiple and complex ways they are actually experienced by communities. These abstractions are made even more intangible by the obscured processes of algorithms, as well as the data they are trained on.



A well-known example was introduced of a system that makes loan eligibility determinations without the use of data relating to protected attributes. What some industry proponents outside of the roundtable saw as an important marker of progress ended up producing biased results just as its predecessors had because data such as savings and continuity of employment were used as a substitute for explicit information on protected attributes. In effect, **the structural nature of biases translates to a legacy that lives differently within data than some might expect**, including in proxy measures that are not viewed as explicitly discriminatory.

Early **citizen education** based around different forms of bias, within and outside of algorithms, contribute to shared societal groundwork in important ways, enabling us to begin laying the foundation for equitable AI even before reaching the stage of trying to build a diverse and innovative team in the industry.

As one participant noted, societal change is also ultimately a matter of individuals continuously advocating for small steps toward equitable AI in their spheres of influence.

How do we help people gain an understanding of the systems we are building (beyond the media)?

Helping the public understand AI systems is a shared responsibility that falls on everyone working in the space. Recognizing that there is no panacea, these efforts are best approached from a variety of angles.

Strategies for the Public Sector

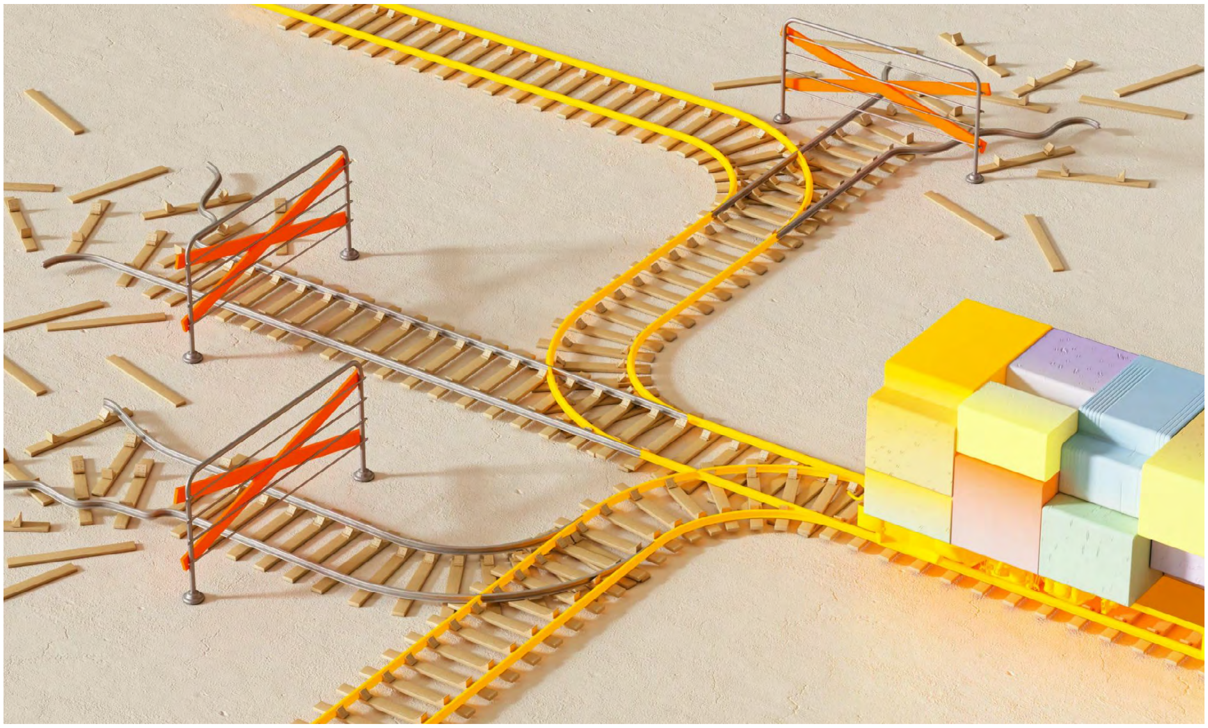
Educational undertakings must begin with an assessment of current AI literacy and the areas where citizens could most benefit from greater understanding or education:

- This knowledge can then be applied toward building freely available massive open online courses (MOOCs) as well as school curricula. Participants in the discussion pointed to successful examples of MOOCs such as Elements of AI, a collaboration between MinnaLearn and the University of Helsinki that has already been translated into a number of European languages.
    - Emphasis should also be placed on AI literacy education at the K-12 level, so as to ensure that it is universal. Further support should then be provided for students continuing these studies in high school, college, and beyond in order to build bridges for society's future leaders. These efforts are particularly important at specific drop-off points that contribute to the diversity and gender gap in computer science.

- Another strategy looks to direct public engagement via town halls and citizens' juries, along with government and privately-supported conversations. The proceedings of these events may then serve as public input for developing legal guidelines for responsible AI.
    - Roundtable-style events that bring together experts across sectors similarly provide an opportunity for information to be fed back to the constituency groups that each of the participants is connected with.
    - Expanding such discussions to a virtual format additionally supports more equitable participation across demographics and geographies. Conversations should be mindful of the public's often-warranted mistrust of big tech and should be viewed as key points of connection in the search to reduce this divide.
- At a more local level, research by one participant's organization has shown that technology use in communities is intermediated, meaning that identifying these intermediaries allows society to know who to support in order to unlock understanding; in many contexts, these figures are youth. One participant recommended that, while remaining cognizant of power dynamics, youth trusted by their communities can be equipped with the education and tools to disseminate meaningful information regarding AI, helping to generate informed discussions and empowered decision making.

Strategies for the Private Sector

Fostering engagement with AI also comes down to the public's ability to navigate the technology. In addition to supporting AI literacy, there are opportunities for **greater investment from companies in interfaces that enable their users**. Systems that include **self-disclosure** and **self-explanation** allow users to distinguish conversations between chatbots and humans, while



also understanding why a given algorithmic recommendation was made to them. Together, these properties facilitate the development of mental models among users.

For instance, the shift from terminal mainframe interfaces to mouses and personal computers was spurred by a small group of researchers at Xerox Palo Alto Research Center (PARC). Similarly, AI developers might ask themselves what the analog of the personal computer is for AI and how to build upon it. After all, the **collective narratives** built about the role of AI in society—particularly its potential as an enabling rather than oppressive force for humanity—influence viewpoints in the push to build this technology differently. Often, the **conversation around AI focuses on harm and extreme circumstances rather than its potential benefits**. While taking care not to downplay the harms of AI, it is important to recognize that fear can push the public away from wanting to learn more.

In addition to more technologically-oriented attempts at creating understanding—including pushing for greater non-technical clarity from the actors developing AI for civil society—there is significant value in furthering off-line education projects. Simply put, people should be able to understand the AI systems the private sector is building without needing to pore over academic papers. Deliberate collaborations with partners such as industry and sector-specific leaders in AI-interfacing industries (medical, education, construction and beyond) allow AI labs to fill gaps in their own approaches, including through alternative channels for creativity. After all, AI systems are the product of more than just factual information—they also involve the imagination. As a result, it is a **natural strategy for labs to deepen their relationships with artists, futurists, authors, and other creative thinkers who can share their visions for what the future should look like**, raise questions they want to see addressed, and propose how to communicate these matters in a language that people can engage with and understand. Though perhaps not a systemic solution, activist efforts in public spaces like buses and restrooms also contribute to efforts of creating understanding.

On the one hand, prioritizing efforts to foster understanding is a matter of resource allocation, meaning that organizations will need to decide who their target audiences will be. For instance, a given AI lab may decide that when allocating resources for their public affairs team, they most value connections with policy makers as well as influencers in the arts and culture community. Regulation also has a role to play in establishing minimum guardrails to ensure affected communities have a path toward justice, even if industry incentives fail to provide this. Therefore, the role of the state and the importance of regulation should not be mutually exclusive from corporate public affairs. It is critical that **policymakers have a deep understanding** of how the AI they are legislating works, as well as the ways it does not work.

What are methods of recourse and reporting when something goes wrong?

At the end of the day, despite challenges in defining what exactly it may look like, experts agreed that **fairness is a basic expectation** for most people. Therefore, as often as explainability and transparency are touted as paths toward progress, participants shared that at a certain level there is no recourse for people who feel that they have been treated unfairly by AI.

Learning From Other High-Impact Sectors

One possible method for working toward recourse and reporting for high impact algorithms draws inspiration from the drug industry. As an example, the U.S. Food and Drug Administration (FDA) has placed increasing responsibility on drug companies over time to prove that their products are safe, effective, and not unduly harmful to certain groups before they are ever able to reach and remain on the market. Similarly, participants suggested that big technology companies should be responsible for ensuring that basic criteria are met and should be held accountable to a regulatory organization such as an 'FDA for algorithms.' Additionally, making the results of testing (such as conditional demographic disparity bias tests) available to the public or to a trusted third-party organization was raised as a means of facilitating accountability.

Supplementing Complaint-Based Recourse

From a legal perspective, AI poses new challenges to recourse, which is put into motion by the raising of a complaint. This is generally sensible for human-to-human interactions, as in the case of anti-discrimination law. Algorithms, on the other hand, complicate this complaint-based system because **people are not aware that they are being wronged or treated unfairly**; the need for this very awareness was raised earlier in the discussion as an addition to the definition of equitable AI.

For instance, based on one's characteristics, the algorithm may tailor search results in ways that hide opportunities such as job postings. These algorithmic decisions take place instantly and out of sight of the searcher, so that the user has no way of knowing what opportunities they are missing out on. This is not to say that complaint-based systems should be abandoned entirely, but instead to suggest the need to think about **new, supplementary regulatory mechanisms** to keep algorithms in check. Adding to the discussion, one participant called on society to think about the ways in which current grievance redressal mechanisms are broken more broadly, including in ways not directly linked to AI.

Where does the burden fall? Policy makers, NGOs, with whom?

It is essential that both the legal and moral burden are not made to fall on either the public sector or the private sector alone. Instead, equitable AI and current failures in achieving it are everyone's responsibility, as is to be expected in a representative republic or democracy. Successful approaches must reflect this shared division of labor in a multi-stakeholder responsibility web or map.

Structural Clarity and Accountability

As participants recognized in the discussion of recourse, fair treatment is at the core of what people seek in their relationship to AI. Achieving this requires a **fuller picture of peoples' experience through greater communication between those on the treatment and outcome sides** of the technology. Taken together, a working understanding of how these facets may or may not fit together can be gained.

While not an answer in itself, asking the very question of where the burden falls serves as a valuable framework for evaluating costs and how they relate to incentives and legal infrastructure. Models, by definition, make mistakes—meaning that costs, failures, and burdens are natural in AI. Sometimes these costs are related to efficiency and the waste of resources, while in other circumstances they may have granular human costs, as in the case of overpoliced communities.

Although AI is often treated as an outlier or an exceptional technology, it is important to recognize that there are **existing frameworks in many democratic systems** governing the implications of technology more generally. Rather than walking away from these frameworks, society can begin from the values of accountability and liability enshrined in protections against the State through constitutions, consumer protection norms, and workers' rights. However, governments and administrations must dedicate additional resources for specific policies related to equitable AI rather than investing disproportionately in general tech policy. Above all, participants agreed on **using law to back up intentions** rather than relying on morality and human goodwill.

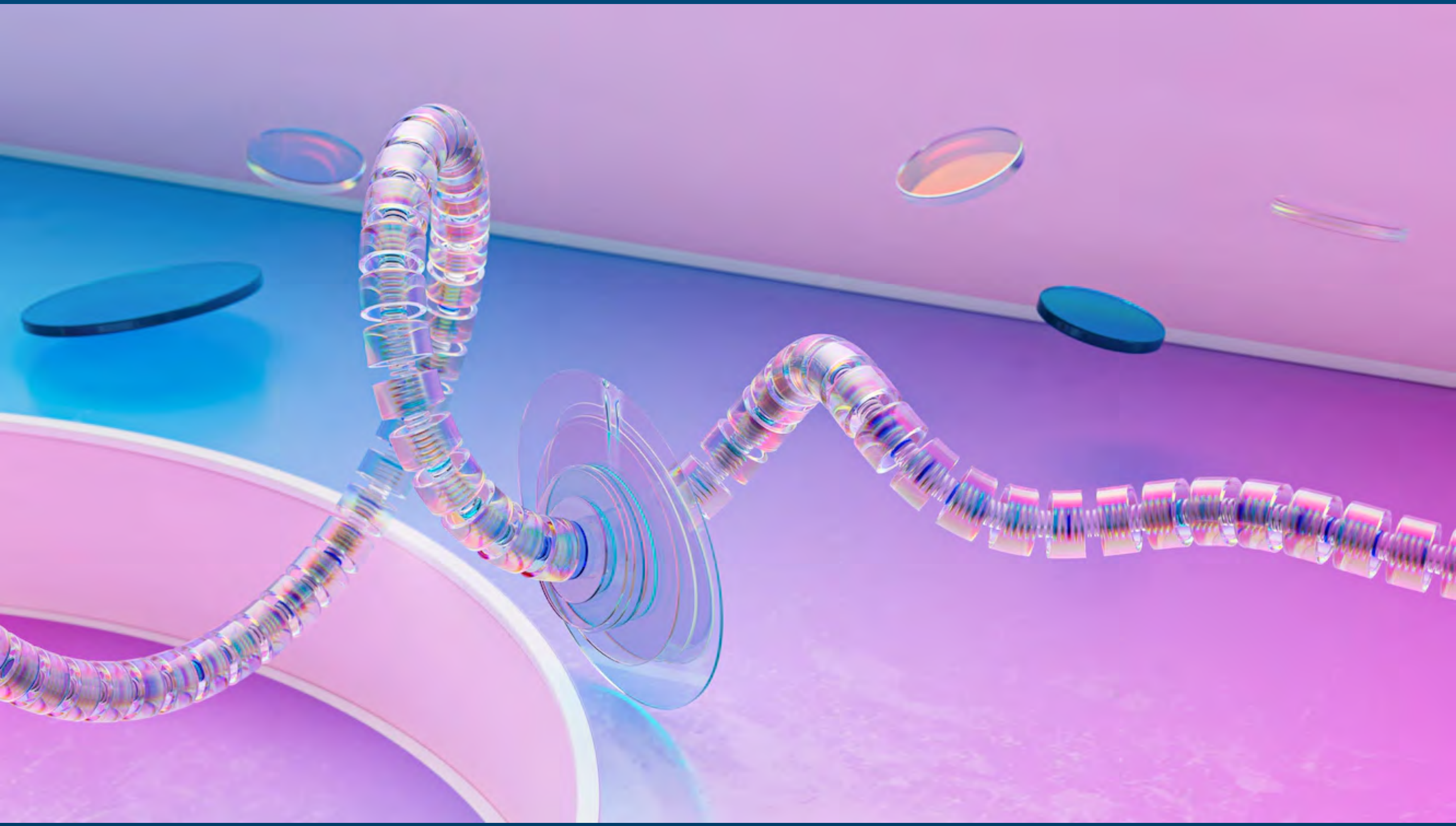
As noted in the preface, this report does not seek to surface issues and ideas that have never been thought of before, nor does it prescribe solutions. Rather, it begins to interrogate what equitable outcomes from AI might look like, explores the paradoxes and promise of the idea, and highlights the value of broad, multidisciplinary conversations. Further discussions in the spirit of those reported here could usefully focus on a) delving more deeply into the questions and challenges reflected in the report, from tightening definitions to aligning incentives, and b) practical ways forward that take account of these challenges.

Appendix: Reading List

Below is a list of **resources** that were shared with roundtable participants before the meetings, as well as resources they recommended after the discussions:

- Johana Bhuiyan, *Vox*: "The Head of Google's Brain Team is More Worried About the Lack of Diversity in Artificial Intelligence than an AI Apocalypse" (08/13/2016).
- Stephen Cave, Kanta Dihal, *Philosophy and Technology*: "The Whiteness of AI" (08/06/2022).
- B Cavello, *The Aspen Institute*: "Making the Case for Trustworthy AI" (09/2022).
- Ulrik Juul Christensen, *The Hill*: "Robotics, AI Put Pressure on K-12 Education to Adapt and Evolve" (09/01/18).
- Flynn Coleman, *Nautilus*: "Who Will Design the Future?" (08/19/2021).
- Iason Gabriel, *Daedalus*: "Toward a Theory of Justice for Artificial Intelligence" (Spring 2022).
- Brent Mittelstadt, *Nature Machine Intelligence*: "Principles alone cannot guarantee ethical AI" (2019).
- Shakir Mohamed, Marie-Therese Png and William Isaac, *DeepMind*: "Decolonial AI" (08/12/2020).
- Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, Alex Hanna, *Patterns*: "Data and its (dis)contents: A survey of dataset development and use in machine learning research" (11/21/2011).
- Inioluwa Deborah Raji, Joy Buolamwini, *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*: "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products" (01/27/2019).
- Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, Parker Barnes, *Proceedings of the 2020 Conference on Fairness, Accountability and Transparency*: "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing" (01/03/2020).
- Inioluwa Deborah Raji, Morgan Klaus Scheuerman, Razvan Amironesei, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*: "You can't sit with us: Exclusionary pedagogy in AI ethics education" (03/03/2021).
- Jackie Snow, *MIT Technology Review*: "We're in a Diversity Crisis: Cofounder of Black in AI on What's Poisoning Algorithms in Our Lives" (02/14/2018).
- Nenad Tomašev, Kevin McKee, Jackie Kay, Shakir Mohamed, *DeepMind*: "Fairness for Unobserved Characteristics" (02/08/2021).

- Sandra Wachter, *European Data Protection Law Review*: "How Fair AI Can Make Us Richer" (2021).
- Sandra Wachter, Brent Mittelstadt, Chris Russell, *West Virginia Law Review*: "Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law" (01/15/2021).
- Sandra Wachter, Brent Mittelstadt, Chris Russell, *Computer Law & Security Review*: "Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI" (03/03/2020).
- Stephen Zorio, *Amazon Science*: "How a paper by three Oxford academics influenced AWS bias and explainability software" (04/01/2021).



SCIENCE & SOCIETY

aspen institute