the CENTER for Civil Rights and Technology

The Leadership Conference on Civil and Human Rights

The Innovation Framework:

A Civil Rights Approach to AI

Acknowledgments

This Framework is intended to move from AI and technology governance principles long established by the civil rights community toward a guide that technology investors, developers, and deployers can reference to ensure that the AI they fund, produce, and use is safe, effective, and fair for all.

The Center for Civil Rights and Technology is a joint project of The Leadership Conference on Civil and Human Rights and The Leadership Conference Education Fund. Launched in September 2023, the Center serves as a hub for advocacy, education, and research at the intersection of civil rights and technology policy. Our experts and partners dive into the most pressing policy issues in three key areas: AI and privacy, industry accountability, and broadband access.

We extend a special thank you to Kasia Chmielinski for their dedication in shepherding this Framework and taking the pen to develop an instruction manual for industry to follow — and for continually pushing us to intentionally build and be creative in our approach.

Staff assistance for this Framework was provided by the Center team, including Koustubh "K.J." Bagchi, Alejandra Montoya-Boyer, Frank Nolan, Frank Torres, Jonathan Walter, and Mariah Wildgen. The Leadership Conference's Christian Madison was the lead designer, Corrine Yu and Patrick McNeil provided proofing assistance, and the Center's Spring 2025 intern Aaron Tiao provided research assistance.

We are grateful for the additional content expertise and feedback that was provided by Jessica Fjeld and Afsaneh Rigot from the De|Center, Tina M. Park, Ph.D., and David Brody.

The author and publisher are solely responsible for the accuracy of statements and interpretations contained in this publication.

Table of Contents

I. Executive Summary

Technology is not innovative if it harms people or leaves anyone behind. True innovation means technology works for all of us, especially communities historically pushed to the margins. People of color, women, LGBTQ+ people, people with disabilities, Limited English Proficient people, older individuals, religious minorities, and immigrants all deserve technology that works for them. This Framework serves as a proactive vision for how emerging technology products, tools, and services, with a focus on AI, can be rights inclusive, safe, and equitable for all people. It addresses how to ensure that emerging technologies can be trusted. Rights-based design and thoughtful creation go hand-in-hand and build a future that's better for all.

Our Innovation Framework includes a series of four core values: 1) Civil and Human Rights by Design; 2) AI Is a Tool, Not the Solution; 3) Humans Are Integral to AI; and 4) Innovation Must Be Sustainable. It also includes 10 pillars along the phases of the AI lifecycle to guide companies as they create and use AI and ensure that their technology protects civil rights, rather than violate them.

The 10 pillars include:

Envisioning Phase

1. Identify Appropriate Use Cases
2. Center Historically Marginalized Users

Design Phase

3. Co-Design with Communities

Training and Development Phase

4. Set Norms and Standards around Build Processes
5. Create and Use Representative Data
6. Protect Sensitive Data
7. Assess for Bias and Discriminatory Impacts

Deployment and Production Phase

8. Close the Feedback Loop
9. Integrate Clear Mechanisms for Accountability
10. Monitor and Improve Consistently

This Framework sets the foundation for what investors and companies must consider when working to invest in, build, and use emerging tech products, tools, and services. The values and pillars are a step toward measuring to what extent companies and investors in specific sectors that utilize consumer-focused tech — such as health care, banking, and housing — are centering impacted communities with a civil rights frame.

Background

II. Background

Who we are

We are advocates and experts who understand that a fair and just society does not differentiate between technological innovation and civil rights. Technology holds the potential to empower community voices, help dissolve historic inequities, and build economic opportunities for everyone, but only with the necessary guardrails to protect against threats, abuses, and bias — both intended and unintended. We deserve technology that works for all of us and understand that equitable technology is also good for business — it's more accurate, more competitive, and more trusted.

Why this Framework and why now

Technology is shaping nearly every aspect of modern life, from work, to education, to health care and other essential services. In the past few years, many emerging technological products, services, and tools have become powered by AI. This Framework comes at a time when governments and businesses are adopting technologies like AI for decision making. While this technological progress can benefit people, many AI tools also carry tremendous risks. These risks are not theoretical. AI-powered systems can work incorrectly due to technical errors, malicious or negligent design, or misuse. Problematic AI-powered systems or uses can result in individuals paying more for the products they buy, failing to be considered for a job, or unfairly paying more for health insurance. Problematic AI systems can deny someone access to public benefits or even falsely accuse them of a crime. These impacts have life-and-liberty-altering and sometimes lethal costs. We have seen these costs occur time and time again. AI is especially harmful when it automates existing biases against marginalized communities, including women, people with disabilities, immigrants, LGBTQ+ people, religious minorities, Limited English Proficient people, older individuals, and communities of color, regardless of the intent of the designers or design. But these real-world harms are not inevitable.

Instead of entrenching flawed AI, reinforcing bias, or automating discrimination — intentional or not — technology must be designed to be safe, effective, and fair for everyone. If AI systems are not safe, effective, or fair, they're not working for any of us. For example, if an AI-powered system is wrongly influencing decisions based on identity traits instead of individual merit, then it is not just unfair — it is also fundamentally inefficient. Such corner-cutting and stereotyping fails to find the most valuable potential customers or employees, and may instead advance less-qualified candidates and lead to bad business decisions. In short, a biased AI system fails at achieving its original intent — neither making business more efficient nor innovative.

We recognize that in addition to policymakers, it is the responsibility of companies and individuals investing in, creating, and using AI and emerging technologies to ensure that the systems they develop and deploy respect people's civil rights. People need assurances that the technology used by companies that make decisions impacting them actually works and works fairly.

Now is the time to move beyond principles towards a future that ensures technologies have appropriate guardrails and center people from the start. Companies that invest in, use, and create AI systems and the people who work at those companies have front-line responsibilities to ensure that goal is achieved.

This Innovation Framework provides a foundation for assessing how industry is incorporating principles such as safety and fairness into the development of their products. We anticipate publishing additional materials to further help guide how the elements of the Framework can be adopted by companies, not unlike the measures taken by companies related to security or data privacy, alongside mechanisms to hold companies accountable.

How we did it

To ensure that this Framework holds true to our values and is applicable and tangible for developers and deployers of AI, we sought a diverse range of input from across the AI ecosystem. We began this process of creating an Innovation Framework in earnest through deep consultation and collaboration with stakeholders in the fall of 2024. In October 2024, we convened a small group of representatives from industry, including leading developers and deployers of AI, along with civil society, to discuss and gather feedback on our initial outline. We also held several small group and individual feedback sessions with members of the civil rights community, the Center's Advisory Council,[i] and individual companies, all of which we used to inform our work.

Our advocacy to ensure that technology works for all of us and the genesis of this Framework started more than a decade ago, with principles and proposed algorithmic and AI safeguards. In 2014, The Leadership Conference, together with leading civil rights and technology organizations, released "Civil Rights Principles for the Era of Big Data,"[ii] which outlined five clear principles: 1) stop high-tech profiling; 2) ensure fairness in algorithmic decisions; 3) preserve constitutional principles; 4) enhance individual control of personal information; and 5) protect people from inaccurate data. In 2020, The Leadership Conference updated these principles[iii] to include ensuring that technology serves everyone; defining responsible use of personal information and enhancing individual rights; and making systems transparent and accountable. These principles, among others, are the bedrock for this Framework.

How to use it

This Innovation Framework turns established rights and principles, agreed upon by civil society and industry alike, into a guide for everyday practice for investors, developers, and deployers of AI. It can be used by C-suite leaders, product teams, and engineers to prioritize effectiveness, fairness, and safety throughout the lifecycle phases of an AI product. It also highlights issues that companies using AI systems should consider before acquiring or deploying that technology. By prioritizing fairness and safety, companies help themselves by building trust with consumers and developing quality products that outlive fleeting trends, leading to sustainable innovation. From this Framework, we will build future tools to help companies implement the values and principles into their processes and standards.

This Framework is also a resource for civil society and others that are looking for guidance on holding companies accountable for the development and deployment of AI systems.

The Innovation Framework®

THE INNOVATION FR

FOUNDATIONAL VALUES

AI DEVELOPMENT PIPELINE

I. ENVISIONING

II. DESIGN

III. TRAINING & DEVELOPMENT

IV. DEPLOYMENT & PRODUCTION

LIFECYCLE PILLARS

RAMEWORK

the CENTER for Civil Rights and Technology

I. CIVIL AND HUMAN RIGHTS BY DESIGN
II. AI IS A TOOL, NOT A SOLUTION
III. HUMANS ARE INTEGRAL TO AI
IV. INNOVATION MUST BE SUSTAINABLE

I. CENTER MARGINALIZED USERS

II. IDENTIFY APPROPRIATE USE CASES

III. CO-DESIGN WITH COMMUNITIES

IV. ASSESS FOR BIAS AND DISCRIMINATORY IMPACTS

V. PROTECT SENSITIVE DATA

VI. CREATE AND USE REPRESENTATIVE DATA

VII. SET NORMS AND STANDARDS AROUND BUILD PROCESSES

VIII. MONITOR AND IMPROVE CONSISTENTLY

IX. INTEGRATE CLEAR MECHANISMS FOR ACCOUNTABILITY

X. CLOSE THE FEEDBACK LOOP

All consumers ought to be able to expect that companies are deliberate in the products and services they envision, design, develop, and use, especially when it comes to AI and emerging technologies. Consumers should be able to trust that these technologies will not harm their lives, their communities, or the world around them. Companies must integrate civil rights and related principles around equity, fairness, and efficiency into core business practices. To advance that goal we have created a Framework of Foundational Values for managing business decisions and specific Lifecycle Pillars aligned with the AI development pipeline to ensure these values are implemented in practice.

Foundational Values

Organizations from individual technology companies and their trade associations, companies using technology created by others, civil society, and governments have all developed ethical or responsible AI principles. These Foundational Values reflect those efforts. With a focus on fairness and trustworthiness, they can help guide corporate decision-makers with managing business decisions in the envisioning, designing, development, and use of AI.

Civil and Human Rights by Design

Every AI system and tool must respect and uphold core tenets of civil and human rights, both in how they are designed and how they are deployed, putting people first. Civil and Human Rights by Design means embedding these rights into every stage of the development process: ensuring that principles like nondiscrimination, privacy, fairness, and accessibility are not afterthoughts, but foundational design requirements embedded into systems. Poorly designed systems can prove costly, and that includes system development that fails to consider civil and human rights at the design stage. Those systems are more likely to create erroneous, less efficient, or unlawful results. When that happens, it harms individuals and communities and it's bad for business. Systems and tools developed to intentionally violate or undermine our civil and human rights cannot be "improved" or made "less harmful" by applying the values and pillars set out in the Innovation Framework and should not be deployed. Examples of systems that should not be built or used are a system optimized for surveillance against Black and Brown communities, or a system that screens out job applicants with disabilities because of a facial tic that has no bearing on an applicant's qualifications.

AI Is a Tool, Not the Solution

Societal issues are complex and deeply entangled with one another. They are influenced by the institutions and policies that govern them and the individual people who are affected by them. AI alone cannot resolve these interlocking issues, but it can serve as one tool among many others to meet challenges and play a role in benefiting everyone. People who are building and deploying AI should recognize this complexity and ensure their AI tools do no harm while, when practicable, improving social, economic, and political equity. This includes working in tandem with broader justice efforts such as policy advocacy, economic community investment, and structural changes that address root causes. For example, an AI-powered algorithm used to help distribute public funding or determine health care delivery cannot address underlying issues in funding levels or care access across all populations if there is a lack of representative data. In this case, the AI system will be limited in its potential impact, or even unintentionally deepen disparities, until other structural changes are prioritized and made.

Humans Are Integral to AI

Understanding human experience is essential to the development and deployment of AI and cannot be fully replaced by AI systems. AI development and deployment must consider the lived experiences of people and communities, including the impacts of AI on these people. AI should complement and uplift human work, not replace or undermine it. For example, organizations ought to consider both the impact of their AI tools on workers and how AI is used within their own operations, such as in efficiency tools and worker management systems. It is also vital to keep a "human in the loop" whenever AI is used in decision making processes. While AI may provide insights, a human should be responsible for making determinations that could impact individuals.

Innovation Must Be Sustainable

New, technology-driven solutions to social problems must be environmentally, socially, and economically sustainable to provide long-term benefits. For example, AI can optimize energy distribution in smart grids to reduce waste, helping communities reduce their carbon footprint and lower energy costs.[iv] It is imperative that the goal of building quickly to gain market share, competitive leverage, and attract customers does not come at the expense of our communities and the building of efficient technologies. The sustainability of AI extends beyond environmental considerations. It also encompasses societal implications and impact on humans, including in the workforce. While AI holds the potential to transform industries and drive economic growth, it must be managed to ensure equitable benefits. For AI to be sustainable, its benefits must be accessible to all.

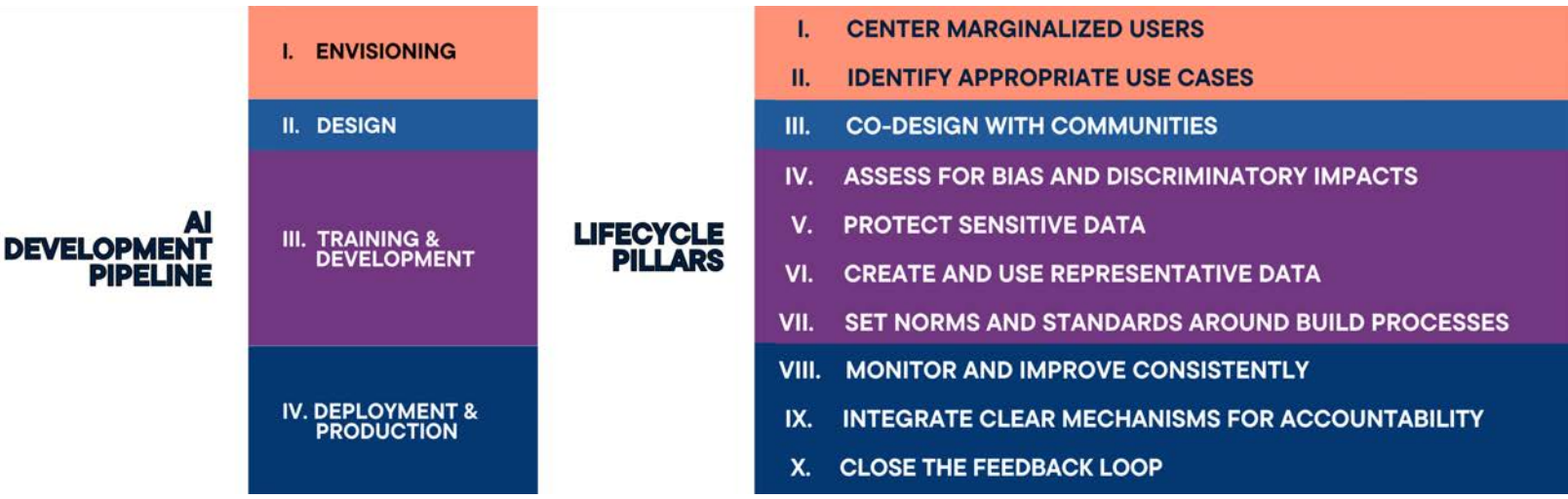
Lifecycle Pillars

Organizations from individual technology companies and their trade associations, companies using technology created by others, civil society, and governments have all developed ethical or responsible AI principles. These Foundational Values reflect those efforts. With a focus on fairness and trustworthiness, they can help guide corporate decision-makers with managing business decisions in the envisioning, designing, development, and use of AI.

Values alone are not enough if they are not put into practice. These Lifecycle Pillars, aligned with the AI development and deployment pipeline, are intended to ensure the foundational values are implemented in practice by C-suite leaders, product teams, and engineers at companies that envision, design, develop, and use AI. Generally, developers control how an AI system is designed and created, while deployers control how an AI system is used and scaled. Both developers and deployers are accountable for AI systems living up to the values and pillars outlined in this Framework.

Development and Deployment: Entities developing or deploying AI systems have shared and distinct responsibilities for maintaining accountability and transparency throughout the AI system's lifecycle, ensuring inclusiveness, and continuously monitoring and improving the system to ensure the AI is transparent, fair, reliable, and secure. Both roles are crucial for the effective use of AI. The exact nature and level of that responsibility demands deeper consideration and discourse with all actors in the AI lifecycle. The Center intends to explore this question in future projects stemming from this Framework.

The AI development pipeline, or AI lifecycle, includes four phases: 1) Envisioning, 2) Design, 3) Training and development, and 4) Deployment and production.



Envisioning Phase

1. Identify Appropriate Use Cases

AI design should be intentional, built with purpose, and implemented with care, recognizing that not all problems can or should be solved with technology. Even those that can be addressed with technology may not be best solved through AI. AI systems should be fit for purpose, with an understanding of how AI is being used in the product or service. Developers should clearly identify the intended uses as well as the capabilities and limitations of their AI systems and should not license systems for off-label uses. Understanding an AI system's capabilities and limitations is critical in assessing whether AI should be used in specific cases. When AI is designed for specific, clearly defined cases, it typically performs better and more reliably than general-purpose AI with broad or open-ended capabilities. Function-specific AI is easier to align with real-world needs, introduces fewer risks of misuse or unintended consequences, and lowers error rates, since it operates within a more constrained and understandable scope.

Additionally, to best inform decision making, it is imperative to explicitly identify the circumstances surrounding end users and people impacted by the technology. For example, a productivity tool that helps an end user organize their calendar or search for published material is very different from a tool to decide where public or private investments should go.

Where AI is not a good technical approach, or where the application of the technology violates civil rights or harms historically marginalized communities, the AI system should not be deployed.

Case Study Using Inaccurate AI to Transcribe Medical Information

Medical AI startups have used generative AI transcription tools to transcribe millions of doctor-patient visits, spanning thousands of clinicians and health systems globally. However, given generative AI's tendency to "hallucinate" and produce factually inaccurate information, this raises serious risks of miscommunication or misdiagnosis. OpenAI's Whisper, for example, has error rates in up to 80 percent of transcriptions,[iv] with nearly 40 percent of inaccuracies deemed harmful or concerning.[v] While AI companies advise against the use of generative AI technologies in "high-risk domains," they do not actively monitor how models are used and deployed. Thus, the importance of developers identifying appropriate use cases is critical to avoiding harm. Some AI use cases, particularly those where errors can lead to real harm, are simply inappropriate for integration into "high-risk domains" like health care systems or law enforcement where an error can lead to physical harm or death.

2. Center Historically Marginalized Users

The internal AI design processes should center historically marginalized users from the beginning of the envisioning process. Centering such users means breaking with the more common methodology of focusing on "ideal" or "normal" users (too often defined as people who share the experiences of those involved in the design process) and considering other users or audiences later. The design stage should explicitly consider the broad array of populations that may interact with or be affected by the AI system and ask whether or how the AI system may operate differently for different people, different contexts, and different geographies. Development, as well as stakeholder engagement, should carefully address specific considerations for historically marginalized communities, as they are at heightened risk of harm from AI models and are often provided the least protection from harm. Many of these harms are already well documented, providing initial insight into the harm that technology can introduce. Importantly, centering historically marginalized users benefits the development process for the entire user base by highlighting gaps, vulnerabilities, and barriers that may not have been identified otherwise. Building a system that addresses the needs and considerations of these communities with respect to privacy, fairness, and accessibility will advance the creation of more efficient and innovative systems that serve customers better, with higher accuracy and potentially lower costs.

Case Study Māori Natural Language Processing (New Zealand)

Building Natural Language Processing (NLP) systems for non-English languages can be difficult because of the lack of training corpus (i.e., the body of data used to train AI) with large volumes of language samples and vetted translations. The aim is most often to expand the number of languages offered by the NLP system, not necessarily to faithfully reflect the nuances of the language's use among its speakers, thereby leading to ineffective models. Te Hiku Media,[vi] a nonprofit established and overseen by indigenous Māori community leaders, aims to preserve te reo, the Māori language, and actively undo violent policies of assimilation that led to the rapid decline in the number of te reo speakers. Te Hiku established new data access and uses indigenous data sovereignty protocols that prioritize Māori values and principles on how the data are used by others. With these data, they build digital tools for language exposure and acquisition not only in close collaboration with living te reo speakers in their communities and with the oversight and consent of local tribes, but also in a way that does not endanger Māori sovereignty.[vii] Thus, the community's data are used to build tools that benefit the community, in line with how the population wants their data to be used. The result is a system that is more accurate and more likely to be accepted and used by its core consumer base. Technology companies building and deploying technology for historically marginalized communities can look to partner with organizations like Te Hiku Media or adopt practices such as custom data governance and usage agreements. In some cases, limited access and use of data may lead to a reduction in product features, which is a tradeoff that should be discussed with those communities.

Resource Design from the Margins Methodology

Companies have used the De|Center's Design from the Margins methodology (DFM)[viii] to design and implement technology and product changes that prioritize the most impacted and vulnerable users. The organization offers practical, step-by-step guidance that is applicable across the stack and on all sizes of project, based on considerations of human rights, justice, and community-based research. Grounded in the knowledge that when those most marginalized are designed for, we are all designed for, DFM allows technologists to identify and mitigate serious harms, designing interventions that benefit all users. Their research, work, and DFM-based changes to technologies has led to features and tools that now affect billions of people but are based on the needs of highly marginalized communities.

The Design Phase

3. Co-Design with Communities

In addition to considering the impact on all people, companies designing AI systems should be proactive in seeking out impacted and historically marginalized communities to build meaningful products that both avoid harm and benefit these communities whose experiences provide priceless input for improved design. Such input can be invaluable to companies developing or seeking to deploy AI systems by identifying both potential concerns and beneficial features. It is important to ensure those who provide input and engage with the company can understand conceptually how the AI system works, what it's intended to do, and how it affects them.

This may require companies to support communities so that they can meaningfully engage during this process, which can include hiring or utilizing community expertise to translate feedback into technical design and building capacity in those communities. Companies should also compensate those consulted for their time and effort. Finally, impacted and historically marginalized communities may include company workers themselves (both employees and contractors), and these populations should also be consulted in the design and build process. If involving historically marginalized communities in a project poses significant risks or potential harm, organizations should seek alternative, intentional ways to incorporate the communities' needs, perspectives, and expertise or methods to mitigate the harm.

Case Study Examining LLM Data for Bias with the Disability Community

A 2023 Google Research study examining biases in Large Language Model (LLM) training data underscores the need to involve impacted and historically marginalized communities in identifying and addressing AI bias.[ix] Researchers worked with people with disabilities to analyze an LLM's outputs, revealing that while responses were rarely overtly offensive, they often reinforced subtle yet harmful stereotypes. Google has used this (and other) research to support their Disability Innovation program, which informs the creation of Google technology that is accessible to people with disabilities.[x] This study highlights the importance of community engagement in uncovering nuanced biases that standard detection methods or external reviewers might miss, creating a better and more trusted product.

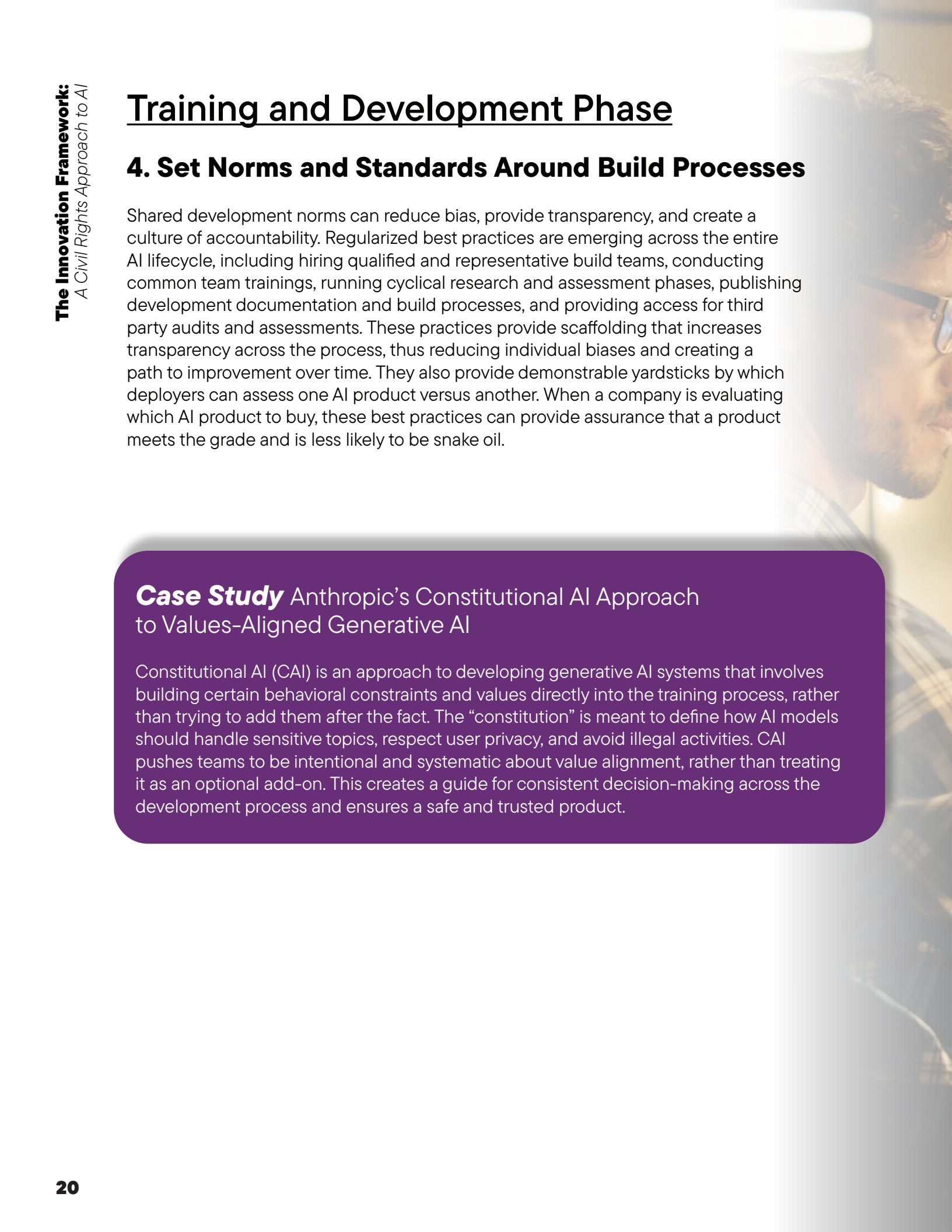
Resource Partnership on AI's Guidance for Inclusive AI

PAI's newly released Guidance for Inclusive AI provides commercial sector AI developers and deployers with a framework for ethically engaging with the public, particularly socially marginalized communities.[xi] Building on the expertise of civil society advocates, industry practitioners, and academic researchers, as well as on insights from different disciplines and domains, the Guidance establishes a different standard for public engagement in the AI sector. It focuses on meaningful public engagement with an emphasis on broadening developers' understanding of the historical and social contexts in which their technology will operate to ensure AI products are designed for the full spectrum of people who will interact with them in their daily lives.

Training and Development Phase

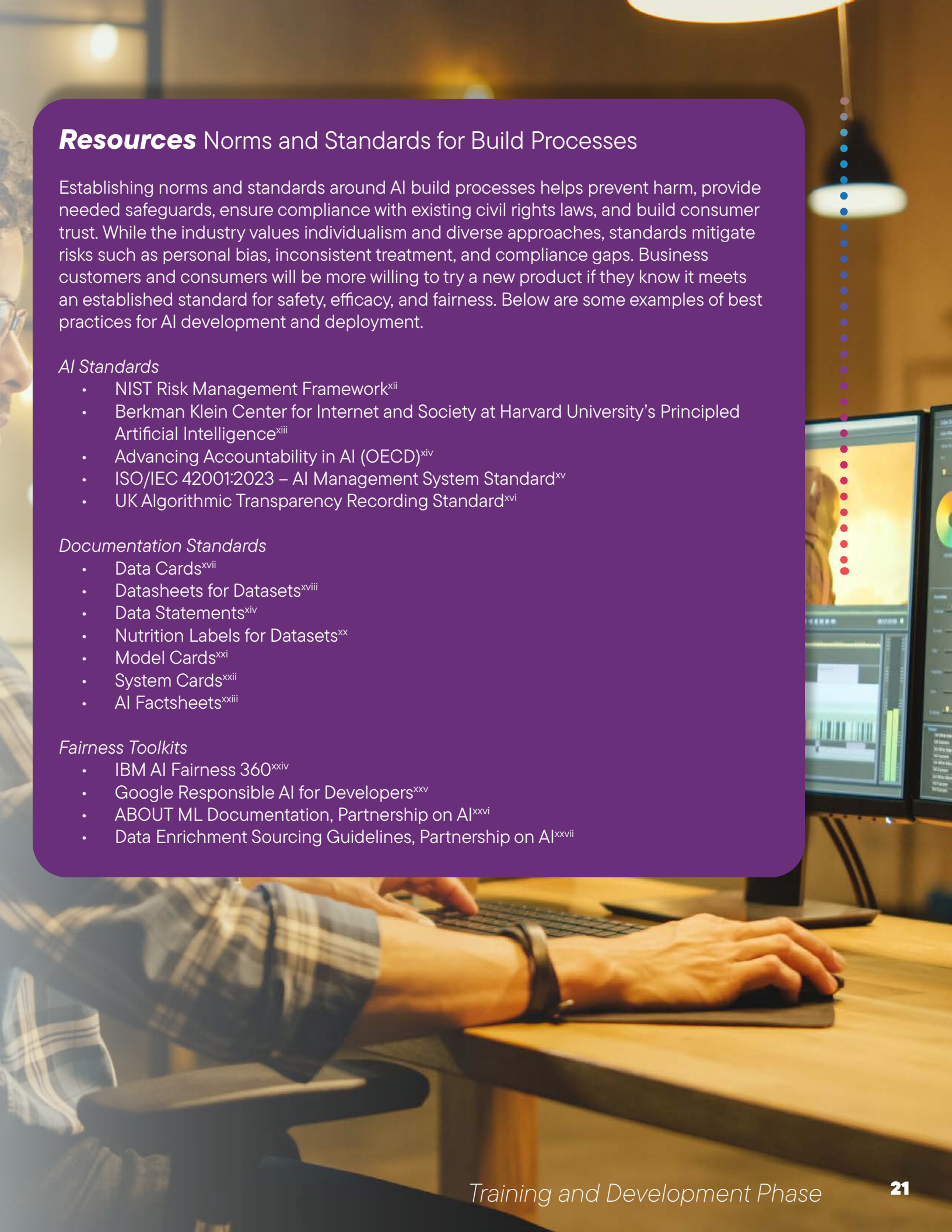
4. Set Norms and Standards Around Build Processes

Shared development norms can reduce bias, provide transparency, and create a culture of accountability. Regularized best practices are emerging across the entire AI lifecycle, including hiring qualified and representative build teams, conducting common team trainings, running cyclical research and assessment phases, publishing development documentation and build processes, and providing access for third party audits and assessments. These practices provide scaffolding that increases transparency across the process, thus reducing individual biases and creating a path to improvement over time. They also provide demonstrable yardsticks by which deployers can assess one AI product versus another. When a company is evaluating which AI product to buy, these best practices can provide assurance that a product meets the grade and is less likely to be snake oil.



Case Study Anthropic's Constitutional AI Approach to Values-Aligned Generative AI

Constitutional AI (CAI) is an approach to developing generative AI systems that involves building certain behavioral constraints and values directly into the training process, rather than trying to add them after the fact. The "constitution" is meant to define how AI models should handle sensitive topics, respect user privacy, and avoid illegal activities. CAI pushes teams to be intentional and systematic about value alignment, rather than treating it as an optional add-on. This creates a guide for consistent decision-making across the development process and ensures a safe and trusted product.



Resources Norms and Standards for Build Processes

Establishing norms and standards around AI build processes helps prevent harm, provide needed safeguards, ensure compliance with existing civil rights laws, and build consumer trust. While the industry values individualism and diverse approaches, standards mitigate risks such as personal bias, inconsistent treatment, and compliance gaps. Business customers and consumers will be more willing to try a new product if they know it meets an established standard for safety, efficacy, and fairness. Below are some examples of best practices for AI development and deployment.

AI Standards

- NIST Risk Management Framework[xii]
- Berkman Klein Center for Internet and Society at Harvard University's Principled Artificial Intelligence[xiii]
- Advancing Accountability in AI (OECD)[xiv]
- ISO/IEC 42001:2023 – AI Management System Standard[xv]
- UK Algorithmic Transparency Recording Standard[xvi]

Documentation Standards

- Data Cards[xvii]
- Datasheets for Datasets[xviii]
- Data Statements[xiv]
- Nutrition Labels for Datasets[xx]
- Model Cards[xxi]
- System Cards[xxii]
- AI Factsheets[xxiii]

Fairness Toolkits

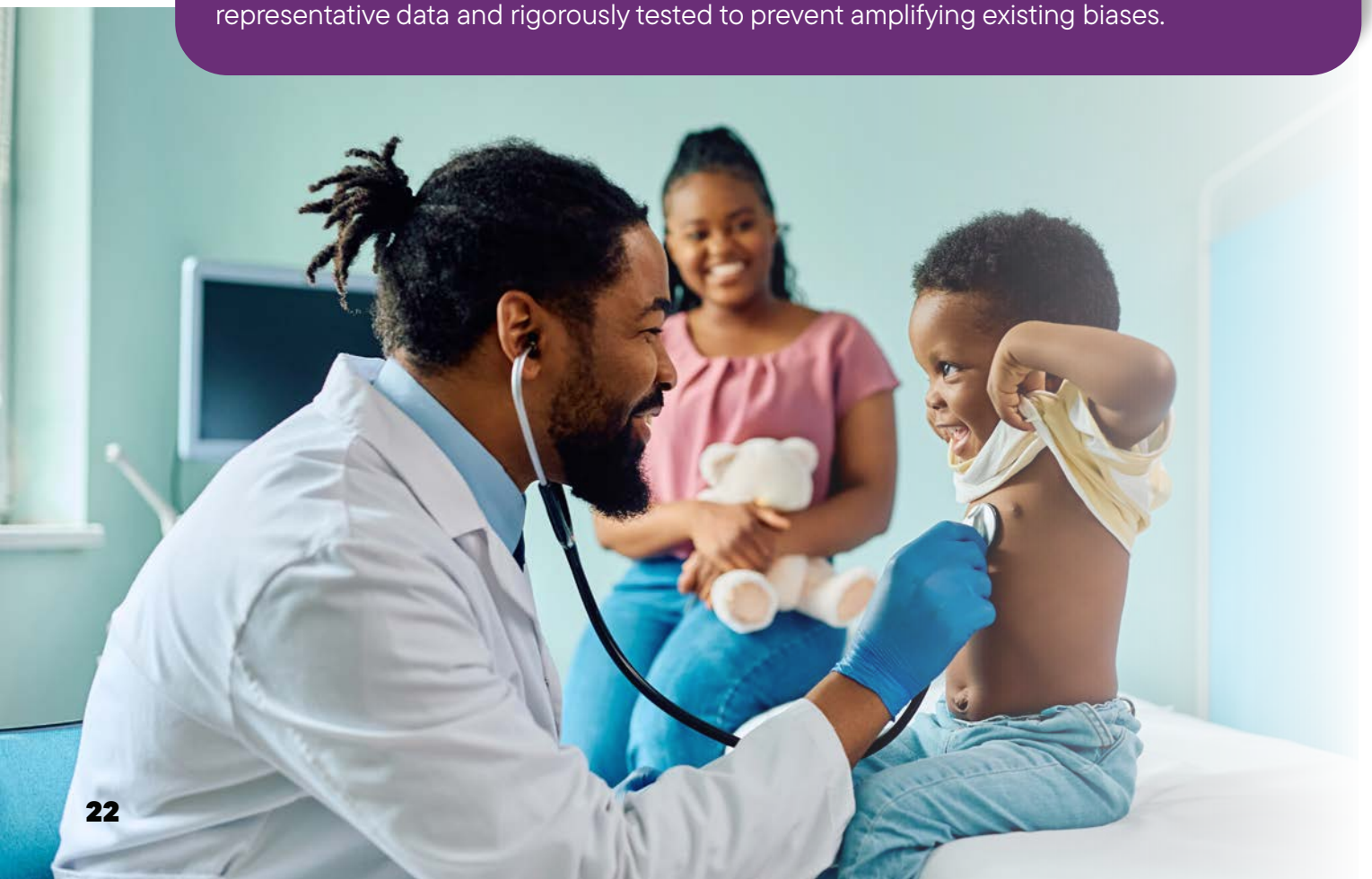
- IBM AI Fairness 360[xxiv]
- Google Responsible AI for Developers[xxv]
- ABOUT ML Documentation, Partnership on AI[xxvi]
- Data Enrichment Sourcing Guidelines, Partnership on AI[xxvii]

5. Create and Use Representative Data

AI should be built on data that are representative and that mitigate foreseeable biases. AI systems will reflect the data used to build the system. If the underlying training datasets include social or historical biases, are under-representative of the communities protected by civil rights laws with whom they will be used, or contain factual inaccuracies, the AI itself is likely to perpetuate those problems. Relevant data include both internal data used for model training and at run-time, as well as third party datasets that are purchased, scraped, or acquired. There is no such thing as a "perfect" dataset for all use cases, so the data used must be mapped and aligned to the expected use case. A one-size-fits-all approach is both ineffective and potentially harmful, both for consumers and the business.

Case Study Unrepresentative Data Causing Algorithmic Bias in Health Care

A health care algorithm used to manage care for 200 million Americans systematically underestimated Black patients' medical needs because it used health care costs as a proxy for health status — overlooking systemic disparities in access to care.[xxviii] As a result, Black patients had to be sicker than white patients to receive the same referrals for critical care programs. Researchers found that adjusting the model to use more representative health indicators reduced bias by 84 percent, highlighting how flawed data assumptions can reinforce systemic inequities. This case underscores the need for AI systems to be built on representative data and rigorously tested to prevent amplifying existing biases.



6. Protect Sensitive Data

Data used to build AI systems or created through AI should be protected and handled carefully. Careless data practices create costly risks to companies and consumers and undermine trust in AI systems and products. The best practice for privacy and data security is data minimization. This means companies should minimize data capture to what is reasonably necessary for a specific purpose, limit data sharing with third parties, and regularly delete stored data that is no longer needed. Data minimization does not mean a company has to minimize absolutely the amount of data it uses; rather it means that data collection and use must be proportional and limited to the need. Some processes (like training a new AI model) may need a lot of data; others do not.

In addition to data minimization, companies should only use private personal data to train AI systems if they have the opt-in consent of the individual to whom the data pertains. People have a right to know and have control over how their data are collected, used, and retained and to challenge consequential decisions made with their data. Beyond informed user consent, companies should also build meaningful user controls that enable people to determine and opt in to what happens with their data. Such personalization and transparency offer multiple benefits: enhancing user security and agency while fostering trust and cultivating sustained product userbases.

Best Practice Privacy by Design

Privacy by design is a proactive product development approach that embeds privacy protections into technology, business practices, and systems from the outset, rather than treating privacy as an afterthought.[xxix] By integrating privacy safeguards throughout the data lifecycle, organizations can enhance user trust, comply with evolving regulations like General Data Protection Regulation (GDPR), and mitigate risks associated with data collection and sharing. Real-world implementations of privacy by design include end-to-end encrypted messaging apps like Signal, search engines that block trackers like DuckDuckGo, systems that provide robust user privacy controls and protective defaults like Apple's iOS, and the Tor browser and network, which ensure browsing anonymity through relays and multi-layered encryption. These examples highlight how privacy by design not only safeguards personal data but also fosters transparency, accountability, and innovation, making it an essential best practice in responsible data management. They also demonstrate how privacy by design can provide a competitive advantage in the marketplace by appealing to consumers looking for more trustworthy platforms.

7. Assess for Bias and Discriminatory Impacts

AI should not automate discrimination or lead to unequal treatment. Before AI systems are deployed, and then as they continue to be operated, maintained, and updated, they must be tested, assessed, and adjusted for unjustified differential impacts, including but not limited to those identified during the co-design phase. Developers should undertake pre-deployment assessments of the design of their algorithmic systems and post-deployment reevaluations based on feedback from deployers. Deployers should undertake pre-deployment assessments of how they intend to use an AI system, and then post-deployment assessments of whether unjustified harm occurred. These types of routinized evaluations can reduce litigation risk to both developers and deployers, weed out unreliable and inefficient technologies, and increase consumer trust. If harm mitigation is not possible, organizations should move to decommission the system. Replication of historic inequities or automation of segregation is not innovative.

Discrimination in AI Disparate Treatment v. Disparate Impact

Discrimination principally occurs in two forms, both of which can manifest or be replicated by AI. First, a system can engage in disparate treatment: It can treat similarly situated individuals differently based on their race, sex, or other protected characteristics as opposed to their individual merit. An example of disparate treatment would be if an AI system down ranked applicants for specific jobs because they are women. Second, a system may cause disparate impacts: This occurs when a system is designed to operate in a facially neutral manner but nonetheless produces results that unfairly and adversely affect a specific group. An example of disparate impact would be an AI system that screens out mortgage or rental applicants with criminal convictions. While facially neutral, such a policy is likely to disproportionately and unjustifiably exclude communities of color that have historically faced unjust policing and overcriminalization, regardless of an individual applicant's ability to satisfy a financial obligation. It is important to note, however, that just because a practice produces a disparate impact or treatment does not mean it will necessarily be illegal or unjustified; there may be reasonable business necessities for a practice that would otherwise be discriminatory. For example, a church using recruiting software to hire clergy could reasonably instruct the system to exclude applicants of other religions because matching the church's religious preferences is a legitimate job qualification. The appendix to the Framework identifies several forms of AI bias.

Case Study Airbnb's Monitoring Tool "Project Lighthouse" Reduces Bias Over Time

Project Lighthouse, developed by Airbnb in 2020, is an assessment tool designed to monitor racial bias on the platform, specifically focusing on "booking success rates."[xxx] Initial racial disparities showed that Black guests booked at a 91.4 percent success rate compared to 94.1 percent for white guests. An updated report in 2023 shows that success rates for all groups rose above 94 percent, though disparities continue to persist. Airbnb continues to invest in monitoring and addressing these disparities through product policies and feature updates. Project Lighthouse demonstrates the importance of ongoing bias assessment to identify and address discrimination during the development and post-deployment stages of a product to ensure fairness for all consumers. Programs like this increase consumer trust in Airbnb's products and services, benefiting the company's bottom line.

Deployment and Production Phase

8. Close the Feedback Loop

Including impacted and historically marginalized communities in the product design is necessary but not sufficient. Developers must also then return to those original designs to confirm they have built in alignment with the original intent, document their actions, and communicate with those communities to close the feedback loop on what was actually built. This should enable conversations about whether the product successfully addressed their concerns or needs to be further modified.

Best Practice Community Consultation Practices at All Stages of Development

Technology companies should involve impacted and historically marginalized communities through the entire process, from design to deployment, ensuring ongoing feedback. While there are consultation programs at some technology companies to engage with external stakeholders early in the design process, these efforts do not represent a complete implementation of this principle, because they very rarely, if ever, encompass closing the feedback loop through re-consultation during the development and post-deployment stages. This is particularly relevant for AI tools built for historically marginalized populations, such as educational tools for children or AI monitoring tools for older individuals, which may be deployed and run in production without any feedback from those communities on whether they address the actual need or have the positive impact initially intended. This gap could be addressed through closing the feedback loop beyond deployment by involving those communities in an assessment of actual value.

9. Integrate Clear Mechanisms for Accountability

For every AI system, there should be clear mechanisms for accountability. Both developers and deployers bear responsibilities as systems are designed, created, and used. This starts with a known responsible and accountable entity (i.e., an organization, other point of contact) that can provide transparency into the inputs and outputs of the system and who can update, pause, or decommission the AI. Further mechanisms can include infrastructure to allow advocates, regulators, or auditors to test AI systems and recourse for individuals who have been harmed by the systems. There must be internal accountability structures and processes, including incorporating civil rights by design into a company's employee training and assessment or auditing processes. It is also vital to have clear lines of responsibility for ensuring those mechanisms are implemented.

Case Study Responsible Practices for Synthetic Media

Clear AI accountability mechanisms — both community-driven and legally mandated — are still emerging. A strong example of a voluntary mechanism is the Partnership on AI's Responsible Practices for Synthetic Media, a framework for the ethical development, creation, and sharing of AI-generated audiovisual content.[xxxi] This initiative brings together industry leaders, researchers, and civil society groups to establish standards that balance innovation with ethical responsibility, ensuring synthetic media benefits society rather than erodes trust. As AI accountability evolves, there is a significant opportunity for industry to lead the way in identifying, building, and communicating new mechanisms for safe and transparent AI.

10. Monitor and Improve Consistently

AI systems should be built to be reliable and stable over time. However, they must also be flexible and responsive to the inevitable shifts of a changing world (such as security issues, user expectations, system bugs, and changes in data input). Developers and deployers should rigorously monitor their AI systems for emerging issues or shifts and update them frequently and regularly to provide better service over time. It is critical that companies evaluate whether an AI system is working properly, without harm or bias. That means documenting and assessing outcomes as well as looking for harmful impacts and inaccurate outputs. Automated monitoring with pre-determined thresholds can also help to rapidly identify and respond to threats or harms, which can reduce the harm footprint to consumers as well as costs to the company.

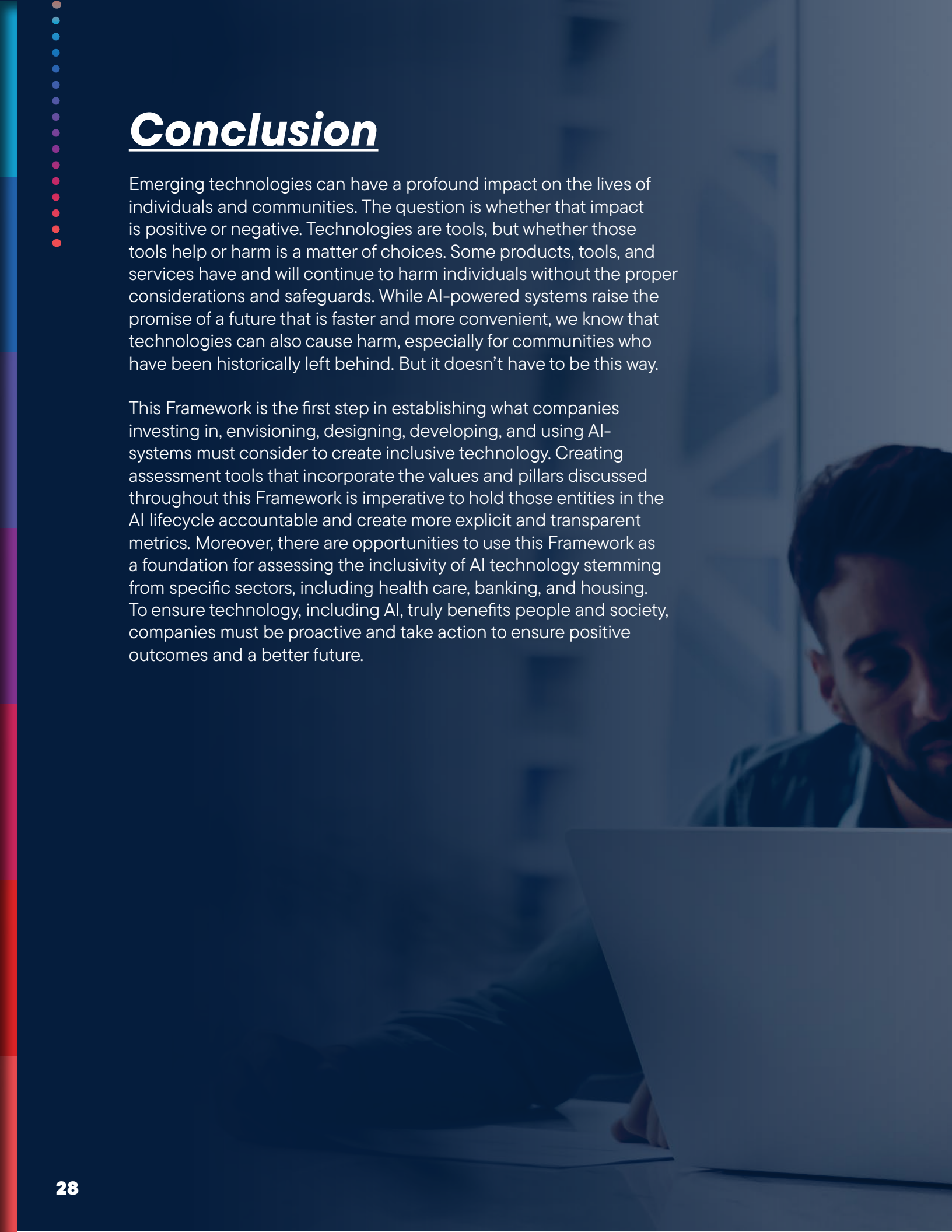
Case Study IDs in Apple Wallet

When Apple launched its "IDs in Wallet" feature, which allows iPhone users to add a digital copy of their state-issued identification card to their Wallet app as an alternative way to provide identity and age verification, it built in mechanisms to monitor for post-deployment bias in the identity verification process.[xxxii] To do so, Apple introduced a voluntary, privacy-preserving data collection mechanism[xxxiii] (differentially private federated statistics), to collect age, sex, race/ethnicity (if available), and apparent skin tone from the identification cards of users who opt-in, in order to determine whether patterns of bias exist between those who successfully complete the verification process to activate their digital ID and those who do not. User data is anonymized and aggregated to minimize any risks associated with sensitive data collection to the user.

Conclusion

Emerging technologies can have a profound impact on the lives of individuals and communities. The question is whether that impact is positive or negative. Technologies are tools, but whether those tools help or harm is a matter of choices. Some products, tools, and services have and will continue to harm individuals without the proper considerations and safeguards. While AI-powered systems raise the promise of a future that is faster and more convenient, we know that technologies can also cause harm, especially for communities who have been historically left behind. But it doesn't have to be this way.

This Framework is the first step in establishing what companies investing in, envisioning, designing, developing, and using AI-systems must consider to create inclusive technology. Creating assessment tools that incorporate the values and pillars discussed throughout this Framework is imperative to hold those entities in the AI lifecycle accountable and create more explicit and transparent metrics. Moreover, there are opportunities to use this Framework as a foundation for assessing the inclusivity of AI technology stemming from specific sectors, including health care, banking, and housing. To ensure technology, including AI, truly benefits people and society, companies must be proactive and take action to ensure positive outcomes and a better future.



Appendix

Forms of AI Bias

There is a growing understanding of different sources and types of bias that can occur related to AI systems. AI bias can happen at different points throughout the AI lifecycle, from the envisioning stage to when a system is deployed and put into use. Recognizing the forms bias can take in AI systems can help to identify, mitigate, and prevent harm. Here is a snapshot of some of the types of bias related to AI, aligned approximately to the relevant phase of the AI lifecycle as defined by the Innovation Framework:

ENVISIONING PHASE

Problem Foundation Bias

Problem Foundation Bias occurs when the basic assumptions underlying a proposed AI system are biased. This can occur when defining the objective, selecting performance metrics, or defining the scope of the problem. Those developing AI systems should examine how problems are defined and framed in the early stages of AI system development to avoid embedding biases. For example, predictive policing systems are especially problematic when based on the assumption that historical crime data are an unbiased representation of actual crime rates. Crime data often reflect policing practices more than actual crime in certain neighborhoods that may be over-policed.

TRAINING AND DEVELOPMENT PHASE

The following types of bias have been aligned to specific aspects of the training and development phase (Data, Features) as indicated below. Please note that in the case of data bias, there is no "unbiased" dataset; there will always be information captured in the dataset and, by definition, data that are left out. Therefore, the goal is not to eliminate all bias but rather to understand bias and address or mitigate the issues that arise regarding misalignment with the dataset to the specific use case (e.g., adjusting the data to include underrepresented populations who are target populations of the systems being built on those data).

Historical Bias (Data)

When AI systems are trained on historical data that reflect past biases, there is a risk of perpetuating or exacerbating those biases. For example, if an AI system designed for use in the criminal justice system is based on historical arrest data, it may disproportionately target certain groups because of systemic and historic biases in law enforcement. If the past was unfair or discriminatory, the AI system will learn to be unfair and discriminatory. Technology is not necessarily a neutral or unbiased arbiter.

Sampling Bias (Data)

Sampling bias occurs when the data used to train an AI system do not accurately represent the population in which the system will be used. When data over- or under-represent certain populations, the system will perform less effectively on those populations. For example, facial recognition systems trained on datasets composed disproportionately of white and male faces perform substantially less accurately when evaluating the faces of women with darker complexions.

Labeling Bias (Data)

Training data are often "labeled" by humans or machines so the algorithm can be trained to identify patterns and correlations. Labeling involves subjective decisions and thus introduces human biases into the AI system; in other words, human judgment is often used to decide what a piece of data is representing. For example, categories and data labels given to images of people could reflect a range of biases. An example would be classifying images of people who look a certain way, such as having several visible tattoos, as "wrongdoers" or "offenders." It is important to have a measure of cultural competency and consider localization in labeling.

Proxy Bias (Features)

Selecting the features or variables that an AI system will use to weigh in making decisions can lead to bias. If an AI system uses a proxy variable to represent an unknowable, sensitive, or otherwise highly complex trait, and this proxy is inaccurate (inconsistently correlated or entirely uncorrelated), it can lead to the creation of systems with unfair outcomes. For example, using ZIP codes as a proxy for wealth in a loan approval can lead to discrimination against people from certain neighborhoods, replicating and reinforcing historic redlining.

DEPLOYMENT AND PRODUCTION PHASE

Confirmation Bias

Confirmation bias happens when we trust information that confirms existing beliefs or ignores information that does not. In the context of AI, a system may be inadvertently designed to predict or return information that confirms preexisting hypotheses by the designer, ignoring contradicting data. For example, human operators may interpret AI results in a way that confirms their own beliefs, further reinforcing bias.

Automation Bias

Automation bias is a form of cognitive bias — a concept that is well-established in psychological research — that occurs when deployers of AI systems defer to the outputs of those systems, even when they contradict the person's own judgment. This is the "the computer said it so it must be correct" bias. Computers make mistakes. This bias is often experienced as overconfidence in the output of algorithmic systems due to the perception that these results are objective, neutral, or necessarily true.

ENDNOTES

[i] The Leadership Conference on Civil and Human Rights. “Center for Civil Rights and Technology Advisory Council.” (last accessed May 7, 2025) https://civilrights.org/ccrt-advisory-council/.

[ii] The Leadership Conference on Civil and Human Rights. “Civil Rights Principles for the Era of Big Data.” Feb. 27, 2014. https://civilrights.org/2014/02/27/civil-rights-principles-era-big-data/.

[iii] The Leadership Conference on Civil and Human Rights. “Civil Rights Leaders Announce Principles to Protect Civil Rights and Technology.” Oct. 21, 2020. https://civilrights.org/2020/10/21/civil-rights-leaders-announce-principles-to-protect-civil-rights-and-technology/.

[iv] Burke, Garance and Schellmann, Hilke. “Researchers say an AI-powered transcription tool used in hospitals invents things no one ever said.” Associated Press. Oct. 26, 2024. https://apnews.com/article/ai-artificial-intelligence-health-business-90020cdf5fa16c79ca2e5b6c4c9bbb14.

[v] Koenecke, Allison; Choi, Anna (Seo Gyeong); Schellmann, Hilke; Sloane, Mona; and Mei, Katelyn X. “Careless Whisper: Speech-to-Text Hallucination Harms.” June 5, 2024. https://facctconference.org/static/papers24/facct24-111.pdf.

[vi] Hao, Karen. “A new vision of artificial intelligence for the people.” MIT Technology Review. April 22, 2022. https://www.technologyreview.com/2022/04/22/1050394/artificial-intelligence-for-the-people/.

[vii] Mahelona, Keoni; Leoni, Gianna; Duncan, Suzanne; and Thompson, Miles. “OpenAI’s Whisper is another case study in Colonisation.” papa reo. Jan 24, 2023. https://blog.papareo.nz/whisper-is-another-case-study-in-colonisation/.

[viii] The De|Center. “What is Design From the Margins?” (last accessed April 16, 2024). https://www.de-center.net/what-is-design-from-the-margins1.

[ix] Gadiraju, Vinitha; Kane, Shaun; Dev, Sunipa; Taylor, Alex; Wang, Ding; Denton, Emily; and Brewer, Robin. “‘I wouldn’t say offensive but...,’: Disability-Centered Perspectives on Large Language Models.” Google Research. June 2023. https://research.google/pubs/i-wouldnt-say-offensive-but-disability-centered-perspectives-on-large-language-models/.

[x] Google Belonging (last accessed April 16, 2025). https://belonging.google/in-products/disability-innovation/.

[xi] Guidance for Inclusive AI (last accessed April 16, 2025). Partnership on AI. https://partnershiponai.org/guidance-for-inclusive-ai/.

[xii] "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile." July 2024. National Institute of Science and Technology. https://www.nist.gov/itl/ai-risk-management-framework.

[xiii] Fjeld, Jessica. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI." Jan 15, 2020. Berkman Klein Center for Internet and Society at Harvard University. https://cyber.harvard.edu/publication/2020/principled-ai.

[xiv] "Advancing Accountability in AI Governing and Managing Risks Throughout The Lifecycle For Trustworthy AI." Feb. 2023. Organisation for Economic Co-operation and Development. https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/02/advancing-accountability-in-ai_753bf8c8/2448f04b-en.pdf.

[xv] "ISO/IEC 42001:2023 – AI Management System Standard." 2023. International Organization for Standardization. https://www.iso.org/standard/81230.html.

[xvi] "Algorithmic Transparency Recording Standard." Nov. 29 2021. GOV.UK. https://www.gov.uk/government/publications/algorithmic-transparency-template.

[xvii] The Data Cards Playbook: A toolkit for transparency in AI dataset documentation (last accessed April 16, 2025). Google. https://sites.research.google/datacardsplaybook/.

[xviii] Gebru, Timnit; Morgenstern, Jamie; Vecchione, Briana; Wortman Vaughan, Jennifer; Wallach, Hanna; Daumé III, Hal; and Crawford, Kate. "Datasheets for Datasets." July 9, 2018. Microsoft. https://www.microsoft.com/en-us/research/uploads/prod/2019/01/1803.09010.pdf.

[xix] Data Statements (last accessed April 16, 2025). Tech Policy Lab. University of Washington. https://techpolicylab.uw.edu/data-statements/.

[xx] The Data Nutrition Project (last accessed April 16, 2025.) https://datanutrition.org/.

[xxi] Model Cards (last accessed April 16, 2025). Google. https://modelcards.withgoogle.com/about.

[xxii] "System Cards, a new resource for understanding how AI systems work." Feb. 23, 2022. Meta. https://ai.meta.com/blog/system-cards-a-new-resource-for-understanding-how-ai-systems-work/.

[xxiii] Using AI Factsheets for AI Governance (last accessed April 16, 2025). IBM Cloud Pak for Data. https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/factsheets-model-inventory.html?context=cpdaas.

[xxiv] Varshney, Kush. "Introducing AI Fairness 360." Sep. 19 2018. IBM. https://research.ibm.com/blog/ai-fairness-360.

[xxv] Introduction to Responsible AI (last accessed April 16, 2025). Google for Developers. https://developers.google.com/machine-learning/guides/intro-responsible-ai.

[xxvi] ABOUT ML Resources Library (last accessed April 16, 2025). Partnership on AI. https://partnershiponai.org/about-ml-resources-library/.

[xxvii] Data Enrichment Sourcing Guidelines (last accessed April 16, 2025. Partnership on AI. https://partnershiponai.org/wp-content/uploads/2022/11/data-enrichment-guidelines.pdf.

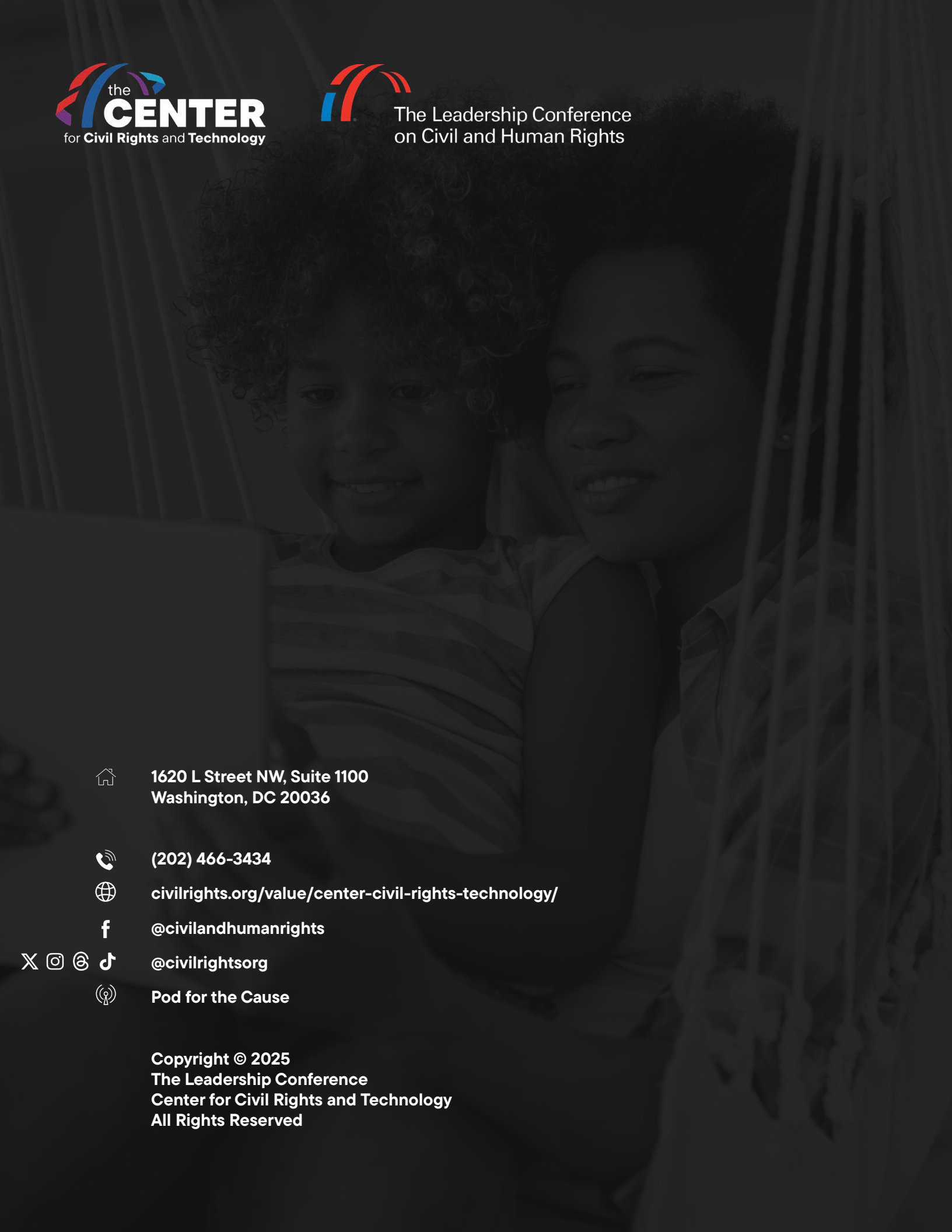
[xxix] A Guide to Privacy by Design (last accessed April 16, 2025). International Association of Privacy Professionals. https://iapp.org/resources/article/a-guide-to-privacy-by-design/.

[xxx] "How we're using data to make travel more open for all." Airbnb. Dec. 12, 2024. https://news.airbnb.com/2024-project-lighthouse-update/.

[xxxi] PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action (last accessed on April 16, 2025). Partnership on AI. https://syntheticmedia.partnershiponai.org/.

[xxxii] IDs in Wallet & Privacy (last accessed April 16, 2025). Apple. https://www.apple.com/legal/privacy/data/en/identity/.

[xxxiii] "Eyes Off My Data: Exploring Differentially Private Federated Statistics To Support Algorithmic Bias Assessments Across Demographic Groups." Dec. 3, 2023. Partnership on AI. https://partnershiponai.org/paper/eyes-off-my-data/10/.



the CENTER for Civil Rights and Technology

The Leadership Conference on Civil and Human Rights

1620 L Street NW, Suite 1100
Washington, DC 20036

(202) 466-3434

civilrights.org/value/center-civil-rights-technology/

@civilandhumanrights

@civilrightsorg

Pod for the Cause