Built to be Fair? Bias, AI Verification and Validation, and the Future of AI

Nancy McCormack

nhou_wei@yahoo.com or McCormack.Wei@mayo.edu

Abstract

Artificial intelligence (AI) and machine-learning (ML) technologies have advanced rapidly and are now deeply embedded across nearly every sector, particularly healthcare. From diagnostic tools and patient-risk assessments to operational decision-making, AI systems influence outcomes that directly affect people's lives. Yet the same data and design choices that give these systems power can also embed hidden, unintentional biases—statistical artifact, historical inequities, or context-blind assumptions that remain invisible until they manifest as unfair or unsafe results.

This paper explores the sources of bias in AI systems and the harm it can cause, particularly in critical fields like healthcare. Using real-world medical examples, it highlights the impact of biased AI and explains the vital role of AI System Verification and Validation (V&V) engineers in detecting, evaluating, and reducing this bias. The paper also presents practical strategies for AI professionals—including AI System V&V engineers—to help ensure that future AI systems are not only powerful, but also fair, trustworthy, and beneficial at both local and global levels.

Biography

Nancy McCormack is a Principal AI/ML Engineer at Mayo Clinic, where she collaborates with data scientists, MLOps engineers, clinicians, and compliance experts to ensure that AI models in healthcare don't just work in production—they work safely, fairly, and reliably for the intended audience group. She also leads the AI Engineering Automation team, building robust automation frameworks, shaping the blueprint for Large Language Models (LLMs), and defining the organization's AI verification and validation (V&V) processes, standards, templates, and best practices.

Before joining Mayo, Nancy spent over 20 years in the tech industry—14 years as a hands-on test and software engineer in the semiconductor, networking, and IT sectors, followed by 8 years in leadership roles managing engineering teams in the semiconductor space.

When she is not busy making AI more trustworthy, you will find Nancy traveling, cheering on sports teams, discovering new food spots with her husband, or binge-watching real-life mystery shows.

She is excited to return to PNSQC as a presenter and paper reviewer—ready to share what she has learned since her last presentation and eager to connect with others who are just as passionate about quality, fairness, and building things that truly make a difference.

1 Introduction

AI systems are now widely used in critical areas like finance, healthcare, law enforcement, and education. As AI makes more decisions that affect people's lives, concerns about fairness, accountability, and bias have become urgent. These issues aren't just theoretical—they cause real harm when AI performs unfairly across different groups.

In healthcare, biased AI can have serious, even life-threatening consequences. AI tools help with diagnosis, risk prediction, treatment, and resource allocation. But many AI models are trained on incomplete or biased data, often missing diverse patient groups, especially from low- and middle-income groups. Bias can also come from the AI algorithms themselves, leading to worse outcomes for minorities, women, older adults, and underserved populations.

Ensuring fairness in AI requires more than good intentions—it demands structured, evidence-based methods. AI System Verification and Validation (V&V) engineers play a key role in detecting and reducing bias throughout the AI lifecycle.

To create fair AI, we must combine ethics with technical rigor—defining fairness clearly, testing across diverse groups, improving transparency, and ensuring accountability. This demands collaboration across disciplines and a strong commitment to inclusive data and global cooperation.

This paper primarily focuses on the medical and healthcare industry and is organized as follows.

- **Section 2: Understanding Bias in AI**
  Discussion of different types of bias including unintentional, with real-world case studies illustrating their consequences.
- **Section 3: AI System Verification and Validation (V&V) as a Framework for Fair AI**
  - How does AI V&V fit into the AI Lifecycle
  - Why AI Verification and Validation are Essential for AI
  - Applying V&V techniques to bias mitigation and model evaluation
- **Section 4: Limitations and Challenges**
  Technical, ethical, and organizational barriers to building fair AI, including issues with data access, measurement of fairness, and accountability.
- **Section 5: The Future of AI: Toward Bias-Resilient Systems**
  - Establishing a global definition of bias and fairness
  - Standardizing and diversifying data
  - Building AI systems both locally and globally

2 Understanding Bias in AI

Bias can arise at various stages of developing AI models and systems. There are several types of bias to consider:

Data bias: underrepresentation, historical prejudice, sampling bias
Example*: A diagnostic model trained and tested mostly on data from urban hospitals may underperform in rural settings where patients present with different conditions or progression patterns. Underrepresentation of minority groups can lead to misdiagnosis or delayed care.*

Algorithmic bias: model behaviors that amplify unfair patterns.

***Example**: A model used to predict who is likely to benefit from follow-up care may favor patients with more documented history—unintentionally favoring wealthier or insured patients who visit clinics more frequently, even if others have greater need.*

 **User/System bias**: interface designs or operational contexts that skew outcomes.

***Example**: A clinical decision support tool might prioritize alerts in a way that assumes all clinicians respond the same, ignoring differences in role or workload. This can lead to critical alerts being missed, especially in busy emergency settings.*

 **User-Pleasing bias**: AI systems are designed to align with user expectations rather than objective outcomes, potentially reinforcing incorrect decisions.

***Example**: A symptom checker might offer "likely" diagnoses that match patient concerns or search history—even when those aren't medically accurate—because doing so increases user engagement.*

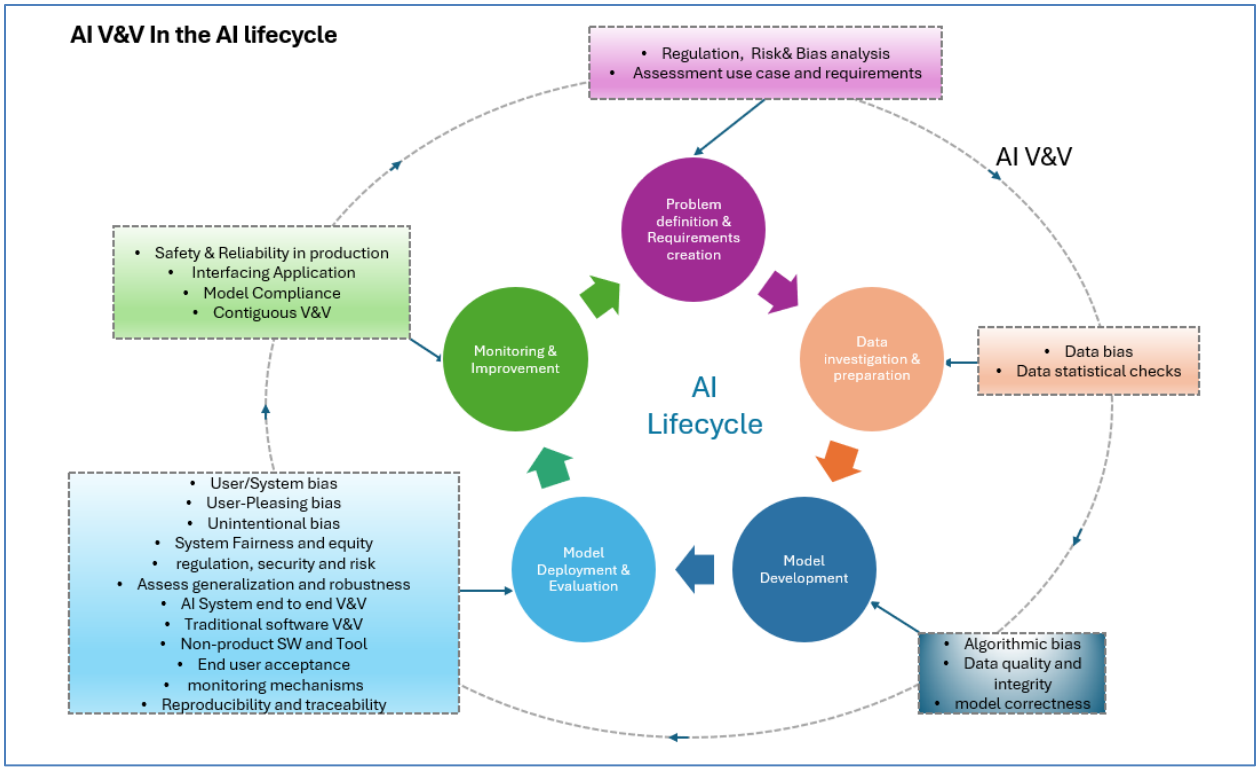
 **Unintentional bias**: Refers to bias that arises inadvertently, often due to unexamined assumptions in model design or deployment.

***Example**: A triage tool that uses zip code as a proxy for health risk might unintentionally discriminate against communities with poor access to healthcare—labeling them low priority due to historical under-utilization.*

3 AI Verification and Validation as a Framework for Fair AI

3.1 AI Verification and Validation as a Framework for Fair AI

- Definitions:
    - ***Verification***: Ensuring the AI system was built right—that is, it conforms to its design specifications and implementation requirements.
    - ***Validation***: Ensuring the right AI system was built, meaning it meets the intended purpose, operates effectively in real-world contexts, and delivers outcomes that are accurate, fair, and trustworthy.



3.2 Why AI Verification and Validation are Essential for AI

AI systems differ fundamentally from traditional software in that they are data-driven, probabilistic, and adaptive. These characteristics make traditional debugging and inspection insufficient. Instead, AI System Verification and Validation (V&V) practices are essential for ensuring that AI systems meet fairness, safety, and accountability requirements and go beyond accuracy: Validating Fairness and Ethics.

Traditional metrics like accuracy or F1 score are not enough. While these metrics help evaluate overall performance, they often mask disparities across different groups or fail to capture ethical and societal risks. For AI systems to be truly fair, especially in sensitive domains like healthcare, they must be validated across multiple dimensions:

Dimension	Description	Example
Demographic Performance Parity	AI should perform equally well across demographic groups (race, gender, age, language, income).	A diagnostic AI tool must not have higher false negative rates for women or minority patients.
Outcome Fairness	AI decisions should align with ethical principles, legal standards, and societal expectations.	An AI system prioritizing organ transplants should not favor patients based on income or location.
Long-Term Impact	Assess fairness not just immediately, but over time—especially for systemic or repeated use.	A triage AI that consistently deprioritizes underserved groups could worsen long-term health disparities.
Data Bias	Bias in training data (e.g., from historical inequities or narrow	An AI model trained only on high-income country data

Dimension	Description	Example
	datasets) directly affects model fairness.	may not generalize well to low- or middle-income populations.
Patient Bias	Variability in how different groups interact with healthcare (e.g., care-seeking, communication) can introduce bias into data and AI outcomes.	Underreporting of symptoms by certain cultural groups may lead to misrepresentation in predictive models.

3.3 Applying V&V to AI Bias

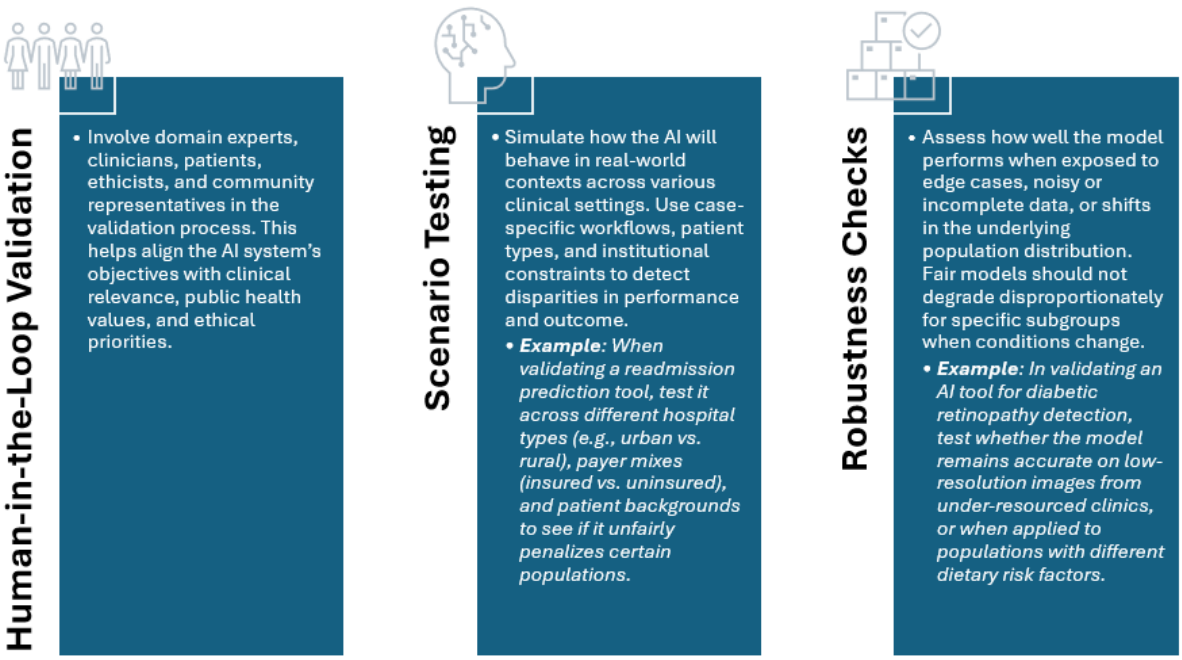
Applying Verification and Validation (V&V) to AI bias means treating fairness and bias mitigation as testable, trackable, and auditable requirements. It involves systematically checking whether the AI system behaves equitably across different subgroups and whether it meets fairness criteria defined for its context. And applying V&V to AI bias in this context ensures that the model's clinical behavior is both safe and fair across diverse patient groups.

3.3.1 Verify Data Diversity and Quality

- **Audit datasets [3]**:
  Representation Across Key Demographics
  Evaluate whether the dataset adequately represents diverse patient populations, including variations in:
  - Race and ethnicity
  - Sex and gender
  - Age groups
  - Comorbidities and health conditions
  - Geographic locations (urban, rural, regional diversity)
  - Socioeconomic status
- **Variations in Data Quality Across Institutions**
  Assess the consistency and reliability of data collected from different hospitals, clinics, or regions:
  - Differences in imaging quality (e.g. older vs newer equipment)
  - Variability in electronic health record (EHR) systems
  - Gaps or inconsistencies in documentation

3.3.2 Validation for Fairness:

Validating AI systems for fairness requires more than just measuring traditional performance metrics. It demands targeted methods that account for social impact, contextual equity, and hidden vulnerabilities. The following strategies support a comprehensive fairness validation process, particularly in high-stakes domains like healthcare:



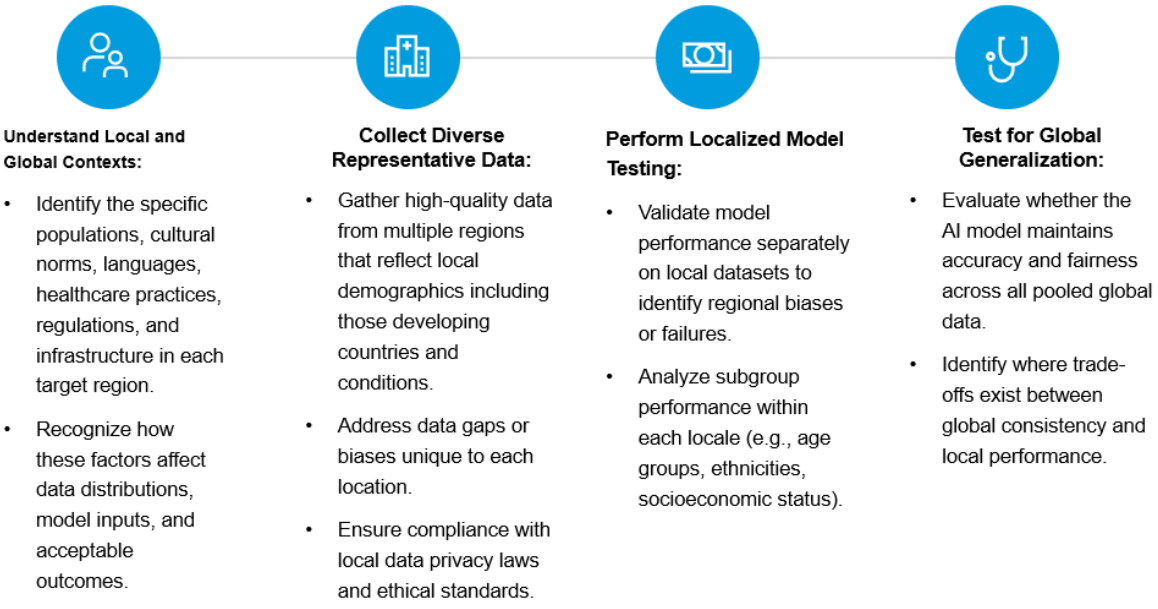
3.3.3 Stress Test on Edge Cases and Rare Conditions

Validate the model's performance on underrepresented or high-risk cases, such as:

- **Rare Disease Patients**
  Validating against rare disease cohorts ensures that predictions remain accurate even when sample sizes are small.
- **Pregnant Patients**
  Pregnancy introduces unique physiological changes that can affect disease presentation, treatment plans, and lab result interpretation.
  ***Example**: A cardiovascular risk model should be validated to ensure it doesn't overestimate or underestimate risk in pregnant women, whose baseline metrics (e.g., heart rate, blood pressure) differ from the general population.*
- **Patients with Multimorbidity (Multiple Chronic Conditions)**
  Many real-world patients have two or more chronic conditions—like diabetes, hypertension, and heart failure—yet many models are trained and validated assuming single-disease cases.
- **Demographically Sparse or Structurally Vulnerable Groups**
  Validate performance on groups that may be underrepresented in training data due to systemic or structural barriers, such as:
  - Non-English speakers
  - Homeless patients
  - Elderly in long-term care
  - Low-literacy or low-health-literacy populations
- **Scenario-Based Testing with Realistic Edge Cases**
  Beyond metrics, simulate testing scenarios using real-world case examples that challenge the system. This includes:
  - Low-resource settings (e.g., limited EHR data or delayed lab results)
  - Noisy or incomplete records
  - Atypical symptom presentations
  - Non-standard clinical pathways (e.g., urgent care instead of ER)

3.3.4 Validating Behavior in Real-World Contexts

- AI systems must be validated not only in training environments but also under real-world operational conditions. Key steps include:
    - **Simulated Feedback Loops**: Test how the AI behaves over time with iterative user feedback to observe potential bias amplification.
    - **Behavioral Monitoring**: Deploy continuous validation tools to monitor shifts in model behavior that may arise from pleasing or aligning with specific groups.
    - **Ethical Alignment Testing**: Go beyond technical performance to ask: Is the model's behavior consistent with ethical, social, and cultural norms—particularly for marginalized groups?
- AI systems must be validated not only locally but also globally.



4 Limitations and Challenges

Despite the strong efforts of AI practitioners to build fair systems with minimal or no bias, there are still significant challenges and limitations that make this goal difficult to fully achieve. These challenges are spanning technical design, data quality, ethical trade-offs, and systemic inequities that make it clear that building fair AI is not a single-step task, but a continuous, collaborative, and multidisciplinary process.

4.1 Technical Challenges

- **Biased or Incomplete Data**
  Historical data often reflects existing social or medical inequalities, and some groups are underrepresented or misrepresented (e.g., minority populations, people with rare diseases).
- **Black-box Models**
  Many modern AI systems, especially deep learning models, large language models (LLMs)
  ***Example**: In LLM-based symptom checkers, subtle language patterns in patient inputs (e.g., describing pain differently across cultures or genders) can lead to biased or inconsistent advice.*
- **Lack of Diverse Testing and Validation**
    - AI is often tested on data similar as what it was trained on.

- Systems may perform well in controlled environments but fail in the real world, especially in under-resourced or global settings.

4.2 Domain-Specific (e.g., Healthcare) Challenges

- **Limited Diverse Clinical Data**
  Many medical AI tools are trained using data from high-income countries or large urban hospitals, leading to bias against rural, low-income, or global populations.
- **Ethical and Legal Constraints on Data Access**
  Collecting sensitive or demographic data (e.g., race, income, sexual orientation) needed for fairness auditing is often restricted by law or privacy concerns (e.g., HIPAA, GDPR).
- **Bias Hidden in Proxy Variables**
  In healthcare, AI models often use proxies like insurance claims or treatment cost—these can reflect systemic inequality rather than actual medical need.

4.3 Social & Institutional Challenges

- **Lack of Standardized Fairness Guidelines**
  There are few universally adopted standards or regulations for measuring or ensuring fairness in AI systems. This leads to inconsistent approaches across organizations and industries.
- **Bias in Human Decision-Making**
  AI systems are trained in human decisions—if those decisions were biased (consciously or not), the model will learn and replicate them.
- **Priority and Resource Constraints**
  - AI development is often driven by speed and performance goals, not fairness.
  - Fairness testing and mitigation require extra time, expertise, and resources, which are often deprioritized.

4.4 Post-Deployment Challenges

- **Fairness Drift Over Time**
  AI models can become biased after deployment as data distributions shift, user behaviors change, or feedback loops reinforce bias (e.g., if underserved patients avoid biased tools, data gets more skewed).
- **Lack of Ongoing Monitoring and Accountability**
  - Fairness is often treated as a one-time evaluation, not a continuous responsibility.
  - Without proper monitoring, biased outcomes can go unnoticed or unaddressed.

4.5 Challenges Introduced by AI Itself

Challenge	Description	Healthcare Examples
Bias Amplification	Small data imbalances can be magnified by AI, especially in deep learning. Models may overlearn patterns that reflect harmful societal biases.	AI associates certain diseases with race/gender due to biased training data.
Reinforcement of Existing Inequities	AI trained on biased historical decisions reproduces and amplifies unfair practices, creating feedback loops.	If past doctors underdiagnosed women for heart attacks, AI might learn and repeat the same behavior.
Overfitting to the Majority	AI models prioritize common patterns and neglect outliers or	Rare diseases or symptoms in minority populations are

Challenge	Description	Healthcare Examples
	minorities, risking poor outcomes for underrepresented groups.	misclassified due to lack of training examples.
Optimization Bias	AI often optimizes for accuracy or efficiency—without fairness constraints, models favor majority outcomes over equitable ones.	A diagnostic tool sacrifices fairness to improve performance on the most common patient demographic.
User Reinforcement Bias	AI systems echo user inputs or behavior, reinforcing biased views, skewed interactions, and short-term preferences over accuracy or ethics.	
• Reinforcing User Beliefs	AI aligns with user biases or assumptions rather than correcting them.	A health chatbot downplays symptoms to avoid user anxiety.
• Rewarding Biased Feedback	AI adapts to user interactions (likes/clicks), which may reflect social bias rather than clinical value.	Triage systems prioritize fast, popular responses over accurate, equitable care.
• Over-Personalization	AI tailoring too much to individual behavior may maintain or worsen existing disparities.	A personalized treatment plan may reinforce unhealthy behaviors or systemic care gaps.
• Ethical Blind Spots	AI avoids unpleasant but necessary information to maintain user comfort or satisfaction.	Systems may avoid telling users about serious conditions due to fear of low satisfaction scores.

5 Future AI: Toward Bias-Resilient Systems

5.1 Establishing a global definition of bias and fairness [1]

As artificial intelligence systems become increasingly embedded in decision-making processes around the world, the need for a unified, global understanding of bias and fairness becomes imperative. Despite the shared goals of promoting equity, transparency, and accountability, existing definitions of these concepts vary significantly across disciplines, cultures, legal systems, and application domains.
Efforts to establish a global definition must account for:

- **Cultural pluralism**: Different societies have distinct norms and historical experiences with marginalization and inequality. What is considered biased in one region may be accepted or even expected in another.
- **Regulatory harmonization**: There is currently a patchwork of regulations (e.g., GDPR in Europe, algorithmic accountability laws in the U.S., ethical AI principles in Asia) that reflect regional values and priorities. Aligning these frameworks will be challenging but necessary for multinational AI systems.
- **Cross-disciplinary input**: Developing a global definition of bias and fairness requires collaboration among ethicists, legal scholars, computer scientists, social scientists, and impacted communities. No single discipline can capture the full complexity of fairness in AI.

5.2 Standardizing and diversifying data

Standardization refers to the development and adoption of consistent protocols for data collection, labeling, annotation, storage, and documentation. Lack of standardization can lead to inconsistencies across datasets, making it difficult to assess fairness, compare model performance, or replicate results.

However, standardization alone is not sufficient. There is also an urgent need to diversify data—to ensure that AI models are trained and evaluated on data that reflects the heterogeneity of the real world. This includes not only demographic diversity (e.g., race, gender, age, ability, geography) but also behavioral, contextual, and linguistic variation. Key strategies include:

Inclusive data source:

Proactively seek data from underrepresented or marginalized groups, rather than relying on convenience samples or historical records that may encode systemic bias.

Bias audits and gap analysis:

Use tools and methodologies to identify over- or under-representation of subgroups and prioritize data augmentation where needed.

Community involvement:

Engage impacted communities in the data collection and labeling process to ensure their values and lived experiences are accurately captured and respected

Localization:

Adapt data collection strategies to local languages, cultures, and norms, especially for AI systems deployed globally or in non-Western contexts.

5.3 Developing Governance for Global Deployment

Developing governance for global AI deployment is critical to ensure accountability, safety, and fairness as AI systems are used across different countries, populations, and legal systems.

- Create policies and processes that oversee AI validation, deployment, and updates across all target regions.
- Ensure accountability and ethical standards are upheld globally and locally

5.4 Building AI Locally and Globally [2]

To create fair, trustworthy, and impactful AI systems, it's essential to strike a balance between local relevance and global scalability. This dual approach ensures that AI systems are both:

- **Responsive** to community-specific needs, and
- **Robust and interoperable** across diverse populations, settings, and infrastructures.

AI systems built solely for global deployment may overlook cultural, economic, and structural differences. Conversely, locally focused systems may lack the scalability and interoperability required for widespread impact. Building AI both locally and globally allows us to maximize equity, efficiency, and inclusiveness.

5.4.1 Engage Local Experts and Stakeholders

- Collaborate with end users and experts like clinicians, regulators, administrators, patients and users in each region, especially people from groups that are often biased against reviewing results and validate relevance and fairness.
- Incorporate their feedback into iterative improvements.

Example: *Mayo Coalition for Heath AI (CHAI), A nonprofit coalition promoting responsible, equitable, and transparent AI in healthcare, with membership across hospitals, regulators, patients, academia, technology vendors, and advocacy groups founding contributions from major health systems and innovators, including Mayo Clinic, Stanford, Johns Hopkins, Microsoft, Google, and others*

CHAI focuses on:

- *Establishing **best practices for AI development, deployment, and oversight***
- *Creating shared **testing and validation frameworks***
- *Promoting **ethical technology adoption that benefits patients and providers***

5.4.2 Document and Report Regional Differences [4]

- Maintain transparency about how the AI performs in different locales.
- Share findings with regulators, users, and development teams to inform updates and governance.

***Example:** In addition to many organizations in the U.S., the Mayo Clinic Center for Digital Health (CDH) has initiated close collaborations with Asian countries—including Singapore and the Philippines—to share best practices, processes, and standards for AI development and implementation.*

5.4.3 Building the world-wide platform for AI

Building a worldwide AI platform is not just a technical challenge, it is a political, ethical, and humanitarian endeavor. Such a platform can help mitigate disparities in access, address global harm before they occur, and foster inclusive innovation that serves all of humanity, not just the most technologically advanced. By emphasizing shared values, coordinated governance, and open collaboration, a global AI platform lays the groundwork for more just and accountable AI systems worldwide.

***Example**: A coalition of nations, led by the UNESCO AI for Good initiative, in collaboration with OECD, IEEE, African Union, EU, and leading universities and nonprofits, launches a project called the Global AI Commons. This platform serves as a collaborative hub for:*

- *Shared **AI policy and ethics standards**,*
- *Open-source **fairness testing tools**, and*
- *Cross-border **data partnerships** to promote transparency and equity.*

6 Conclusion

AI systems hold transformative potential, but they also risk introducing or deepening biases—especially when models seek to align with user preferences or societal norms without sufficient oversight. These risks become especially salient in global, real-time deployments where emergent behaviors can reinforce inequality. Robust verification and validation practices, rooted in diverse data, ethical oversight, and global perspectives, are essential to guide the development of AI that is not only powerful and effective, but fundamentally fair.

Bias in AI is not only a technical issue, but a reflection of societal structures and design choices. As AI systems become more powerful and widespread, they must be held to higher standards of fairness, accountability, and trust. Verification and Validation (V&V) offer a rigorous framework to assess these dimensions. By embedding V&V throughout the AI lifecycle—from design and data collection to deployment and monitoring—we can move toward systems that are not only intelligent but also just. The future of AI fairness depends on whether we take V&V seriously today.

Acknowledgements

Thank **for Pallavi Sharma, Tafline Ramos** and **Sam Simataa** their valuable comments and feedback on earlier drafts of this paper.

References

[1] Leo Anthony Celi, *Massachusetts Institute of Technology, Beth Israel Deaconess Medical Center, Harvard T.H. Chan School of Public Health,* "From AI Bias to AI by Us". Available:

https://pmac-2025.com/local/storage/uploads/sessionMaterial/Leo%20Anthony%20Celi-20250221-315496.pdf

[2] Leo Anthony Celi, *Massachusetts Institute of Technology, Beth Israel Deaconess Medical Center, Harvard T.H. Chan School of Public Health,* "Joining the battle against health care bias". Available:

https://news.mit.edu/2023/joining-battle-against-health-care-bias-leo-anthony-celi-0516

[3] Peter A. Noseworthy, MD, Zachi I. Attia, MSc, LaPrincess C. Brewer, MD, MPH, Sharonne N. Hayes, MD, Xiaoxi Yao, PhD, Suraj Kapa, MD, Paul A. Friedman, MD, and Francisco Lopez-Jimenez, MD, MSc, "Assessing and Mitigating Bias in Medical Artificial Intelligence: The Effects of Race and Ethnicity on a Deep Learning Model for ECG Analysis". Available:

https://www.ahajournals.org/doi/10.1161/CIRCEP.119.007988

[4] Overgaard, Shauna M., Ph.D. Gai, Chenyu Ohde, Joshua W., Ph.D, "Guiding responsible AI in healthcare in the Philippines". Available:

https://www.nature.com/articles/s41746-025-01755-3