WSIS FORUM 2026

6–10 JULY 2026 · GENEVA ITU

Session Outcome Document

Fit for Whom? Sex-Stratified Data and the Integrity of High-Risk AI

Women At The Table / FemTechnology

Tuesday, 7 July 2026 · 10:00–10:45 · Room H2, ITU Montbrillant, Geneva

WSIS Forum 2026 — Session 184

Key Issues discussed: Looking Beyond 2025

- The session examined whether high-risk AI systems can be considered valid, trustworthy, or fit for purpose when they are evaluated primarily through aggregate performance metrics. Speakers highlighted that average performance can conceal systematic failures for women and other underrepresented groups, particularly in domains such as health, justice, employment, welfare, and public administration.
- The discussion reframed sex-stratified data as a question of data integrity, clinical validity, and public accountability, not only as a gender-equality concern. If an AI system affects women but cannot demonstrate how it performs for women, it should not be described as validated for women.
- **Speakers discussed how women can disappear before AI enters the picture:** through incomplete clinical evidence, male-default diagnostic norms, missing life-stage variables, biased institutional records, and unstructured data that cannot be tested or audited.
- Health was discussed as a particularly clear example of the problem because it shows how historical evidence gaps can become automated decision-making risks. Women's symptoms have often been under-recognized, pain underdocumented, adverse events underreported, and sex-specific disease presentation insufficiently reflected in clinical evidence, diagnostic pathways, and structured health records. The discussion therefore highlighted the importance of structured, clinically relevant data: such as sex, reproductive life stages (such as postpartum or menopause etc.) which may all affect whether a system is valid for its intended use.
- Speakers emphasized that model-level fairness interventions cannot fully repair upstream data failures. A model cannot recover what the record never captured. AI trained on incomplete institutional evidence may reproduce those omissions at greater speed and scale while appearing technically efficient or neutral.
- The session connected health data gaps to broader AI governance challenges across public systems. Similar risks arise wherever institutional records have historically underrepresented, misclassified, or misunderstood women's experiences, including justice systems, labour markets, public benefits, migration systems, and digital public infrastructure.
- The discussion brought in a development and public-sector lens, including the importance of country capacity, institutional readiness, procurement, standards, and digital public infrastructure. Governments need the ability to evaluate AI systems before deployment, ask whose data is represented, require disaggregated evidence, assess local relevance, and monitor harms after systems are in use.
- Participants discussed the particular risks of cross-border AI deployment. Systems validated in one country, institution, language, or population should not be presumed safe or effective in another without local-context assessment. This is especially important for developing countries and Global South contexts that may be asked to adopt systems trained or validated elsewhere.
- The session situated sex-stratified validation within the broader responsible AI and sustainable development agenda. Trustworthy AI should support human rights, gender equality, digital inclusion, public value, and institutional trust. These goals must be built into the design, procurement, deployment, and oversight of digital systems, rather than added after harm occurs.

Tangible Outcomes of the session

- **Speakers highlighted a core governance question for AI systems:** fit for whom? This question asks whether a system has been tested only on average, or whether it can demonstrate how it performs for the people and communities subject to its outputs. The session clarified that sex-stratified data and disaggregated performance reporting are not narrow technical details. They are central to validity, safety, accountability, and trust in high-risk AI systems.
- The health discussion made the issue concrete. Speakers highlighted how gaps in sex-specific evidence can affect diagnosis, triage, treatment recommendations, drug safety, referral pathways, symptom interpretation, and risk prediction. When these gaps are built into training data, AI can make inherited blind spots faster, cheaper, and harder to see.

- The session showed how health AI, justice AI and other public-sector systems can share a common failure pattern: incomplete upstream records can produce downstream systems that appear accurate overall while failing specific groups.
- The discussion emphasized that AI governance must account for the institutions and data systems beneath the model. In health, this means examining whether the data record captures the realities needed to evaluate women accurately, including sex-specific presentation, life-stage context, treatment response, adverse events, and diagnostic delay where relevant.
- The discussion emphasized that AI readiness is not only about infrastructure, innovation, or adoption. It also includes the capacity of public institutions to evaluate data quality, assess subgroup performance, govern vendor relationships, protect rights, and respond when systems fail.
- The session reinforced the importance of local validation and contextual assessment. High-risk AI systems should be evaluated in relation to the legal, institutional, linguistic, demographic, epidemiological, and social realities of the setting where they will be used.

Key Recommendations and Forward-Looking Action Plan for the WSIS+20 Review and Beyond

- High-risk AI systems should not be certified, procured, or deployed solely on the basis of aggregate performance. Where systems affect women and girls, deployers should be required to show how the system performs by sex and by other relevant characteristics where lawful, appropriate, and material.
- Sex-stratified validation should be treated as a minimum condition of validity in high-risk contexts. If a system cannot demonstrate how it performs for women, it should not be described as validated for women.
- Public authorities should embed disaggregated validation requirements into procurement templates, AI impact assessments, public-sector AI registries, audit protocols, technical standards, and digital public infrastructure governance.
- In health AI, systems should document the sex distribution of training, validation, and deployment populations; assess performance by sex; disclose missing or inconsistent data; and capture clinically relevant variables where they are material to safety, effectiveness, or validity.
- Clinically relevant health variables should be structured and testable where they affect the system’s purpose. These may include sex, pregnancy status, reproductive life stage, menopause status, hormone exposure, sex-specific symptom presentation, pain patterns, medication response, adverse events, and referral or diagnostic delay.
- Missing subgroup data should be treated as a governance and safety risk. Public institutions should ask not only whether a model is biased, but what the system could not know because the institution never captured, structured, or measured it.
- Countries and public institutions should require local-context assessment before deploying high-risk AI systems developed or validated elsewhere. No system should be presumed valid across populations, languages, legal systems, health systems, or institutional contexts without evidence.
- Developing countries and Global South governments should be supported in building the capacity to evaluate AI systems, negotiate procurement requirements, demand transparent evidence, and refuse systems that cannot demonstrate performance for affected populations.
- Digital public infrastructure should be designed so that gender equality, human rights, inclusion, and accountability are built into data standards, governance processes, procurement, and oversight from the beginning.

Suggested Follow-Up / Input to WSIS 2035 and AI Governance Processes

As AI systems become embedded in health systems, justice systems, labour markets, welfare systems, public administration, and digital public infrastructure, governance must move beyond broad principles and average metrics. The next phase of AI governance should focus on whether systems can demonstrate performance across the populations they affect.

Health shows the stakes clearly. If women’s symptoms, life stages, adverse events, medication responses, pain patterns, and disease presentations are missing or poorly structured in the underlying record, AI will not correct the evidence gap. It may scale it.

WSIS follow-up, the Global Digital Compact, the Global Dialogue on AI Governance, national AI strategies, technical standards, and public procurement processes should treat sex-stratified validation and disaggregated performance reporting as baseline requirements for high-risk AI.

“Fit for purpose” cannot mean fit on average. It must mean fit for the people, institutions, and contexts where AI systems are deployed.

Related FemTechnology research and implementation resources

- Invisible by Design — Women’s Health as the Blind Spot in AI and Medicine
- The Architecture of Women’s Health — Outcomes & Action Framework
- FemTechnology Research & Intelligence