WOMEN AT THE TABLE

AI & EQUALITY

We Shaped Our Tools. Now They Are Shaping Who Lives, Who Is Believed, and Who Is Left Behind.

AI Cannot Outperform the Evidence It Was Built On

Women At The Table | 2026

The Thesis, Revisited

In 2019, Women At The Table published *We Shape Our Tools, Thereafter Our Tools Shape Us* — a position paper warning that automated decision-making systems trained on biased data would not correct inequality but cement it. We argued that machine learning makes the implicit information in data explicit in the code, and then "intelligently mirrors" the information it has been given from the analog world. We called for Affirmative Action for Algorithms: deliberate intervention to ensure that the biases being slowly stripped from society were not being hardwired into the digital infrastructure of the century ahead.

Six years later, the evidence is in. The warning was not speculative. It was predictive.

Two of the systems most consequential to women's lives — healthcare and criminal justice — are now being automated at scale. In both, the same structural pattern is

repeating: historical exclusion of women from the foundational evidence base produces biased data; biased data trains biased algorithms; biased algorithms generate biased outcomes; and those outcomes are fed back into the system as new "ground truth." The loop closes. The bias locks in.

What has changed since 2019 is not the direction of travel but its velocity. Healthcare is adopting AI 2.2 times faster than the rest of the economy. Courts across at least twenty U.S. states, multiple Canadian provinces, several Australian territories, and pilot programs across Europe now use algorithmic risk assessment tools for bail, sentencing, and case management. The window to intervene before these systems calcify is narrowing.

This paper traces the common architecture of bias across both domains, draws on Women At The Table's recent research — *Invisible by Design*: *Women's Health as the Blind Spot in AI and Medicine* and *Gender Bias in Judicial Algorithms*: *A Global Analysis of Algorithmic Discrimination* — and sets out an integrated agenda for action.

I. The Same Flaw, Two Systems

The foundational problem in both healthcare AI and judicial algorithms is identical: the systems that generated the training data were never designed with women in mind.

In Medicine

Clinical research defaulted to male physiology for most of the twentieth century. The U.S. FDA banned women of childbearing potential from early-phase trials until 1993. Female representation in randomized controlled trials still averages roughly 37%, with three-quarters of studies reporting no sex-stratified outcomes. In neuroscience, single-sex animal studies using males outnumber those using females by approximately 5.5 to 1. The preclinical pipeline — cell cultures, animal models, pharmacokinetic studies — was built around the male body.

From this foundation, clinical guidelines, diagnostic thresholds, and reference ranges were calibrated on male physiology. Standard cardiac troponin thresholds miss heart attacks in women. The MELD score for liver transplant prioritization systematically underestimates disease severity in female patients. Drug dosing protocols derived from male-dominated trials produce adverse reactions in women at nearly twice the rate observed in men. When electronic health records encode these guidelines as structured data, and when AI models train on those records, the male default becomes the algorithmic default.

In Criminal Justice

Judicial algorithms learn from historical court decisions shaped by decades of gender stereotypes. In Brazil, BERT language models detect gender bias in court decisions with 88.86% accuracy, revealing systematic patterns where women are characterized through emotional language while men's actions are described through situational factors. When those linguistic patterns become training data, the algorithm learns to associate women with unpredictability and men with circumstantial violence.

Even when gender is removed as an explicit variable, algorithms discriminate through proxies. Employment gaps penalize women with caregiving responsibilities. Residential instability scores reflect housing inequality. Relationship history variables encode assumptions about dependence and autonomy. The algorithm never names gender — and gender is everywhere in it.

In both domains, the same dynamic applies: **the system was built on evidence that treated the male experience as universal, and the algorithm faithfully reproduces that assumption at scale.**

II. The Feedback Loop: How Bias Becomes Infrastructure

Our 2019 paper identified the core mechanism: machine learning absorbs human bias, replicates it, incorporates it into future decisions, and makes implicit bias an explicit reality. In both healthcare and justice, this mechanism has matured into a self-reinforcing loop.

In Healthcare: The Cascade of Distortion

The bias cascade runs through six layers. Male-centric clinical research (Layer 1) produces male-centric guidelines and diagnostic thresholds (Layer 2). Clinicians trained on those guidelines document what they see through the same lens (Layer 3). Their documentation populates electronic health records that become the de facto ground truth for AI (Layer 4). Language models and clinical AI tools train on this biased corpus (Layer 5). And the resulting algorithms produce outcomes — delayed diagnoses, missed conditions, undertreated disease — that compound in real lives (Layer 6).

The endometriosis case illustrates the full loop. Women wait an average of seven years for diagnosis. During those years, their records accumulate incorrect codes — pelvic pain, irritable bowel syndrome, anxiety — rather than the actual disease. When AI trains on those records, it learns the pattern of human error: late-stage disease, not the early signal that could have prompted intervention. The model does not correct the delay. It reproduces it.

In Criminal Justice: The Discrimination Cycle

The feedback loop in judicial algorithms follows the same architecture. Historical arrest data reflects policing patterns shaped by racial and gender bias. Court records embed credibility assessments where women's testimony is systematically devalued. Sentencing data reflects disparities in how violence by and against women is categorized and punished. When algorithms train on this data, they learn the priorities and blind spots of the systems that generated it.

The COMPAS system demonstrates the result: women rated "high risk" had less than half the actual reoffending rate of men rated "high risk" — 25% versus 52%. The algorithm systematically overpredicts women's risk. In the Netherlands, the HART

system rated Moroccan and Turkish immigrant women 40% higher risk than Dutch women with identical criminal histories. In Australia, Aboriginal women were systematically rated higher risk despite lower actual reoffending rates.

In both domains, the loop closes the same way: **biased outputs become new training data, and the system's errors become its assumptions.**

III. The Patterns Between

Examining these two systems side by side reveals structural parallels that illuminate why technical fixes alone cannot solve the problem.

1. The Credibility Gap, Automated

In healthcare, women's symptoms are more likely to be dismissed as psychological, labeled as stress or anxiety rather than investigated as physical illness. A population study of 6.9 million patients across 770 diseases found women are diagnosed an average of four years later than men — not because disease appears later, but because early symptoms are minimized.

In criminal justice, women already face significantly lower credibility ratings from both male and female judges, particularly in sexual assault cases. When algorithms train on historical decisions embedding these credibility biases, they systematize the gap. Natural language processing systems learn to associate women's speech patterns with uncertainty. Sentiment analysis tools interpret women's emotional expressions as less reliable.

In both systems, what begins as human skepticism toward women's accounts becomes machine-level certainty. **The algorithm does not doubt women. It has simply never learned to believe them.**

2. The Impossibility of Neutral Algorithms

Computer scientists have proven mathematically that no algorithm can simultaneously achieve predictive parity, equal false-positive rates, and equal false-negative rates when base rates differ between groups. This is the impossibility theorem — not a limitation to be engineered away but a mathematical proof.

In healthcare, this manifests as models that perform well on aggregate metrics while systematically underperforming for women. Cardiac disease prediction algorithms trained on data balanced equally between men and women were still consistently less accurate for women across sixteen independent experiments: false-negative rates were higher for female patients even with equivalent data volume.

In criminal justice, twenty-one different mathematical definitions of fairness cannot all be satisfied simultaneously. Choices must be made — fair to whom, measured how, at whose expense. These are political decisions wearing technical masks.

In both domains, the lesson is the same: **fairness is not a parameter to be optimized. It is a value that must be designed in, demanded, and enforced.**

3. The Missing Variables Problem

Both systems suffer from a shared structural absence: they cannot see what was never recorded.

In healthcare, most EHRs lack structured fields for menstrual cycle phase, menopausal status, pregnancy history, hormonal therapy, or postpartum recovery. These variables profoundly alter the clinical significance of symptoms and lab values. An algorithm predicting cardiovascular risk for a 45-year-old woman cannot factor in that she is perimenopausal — because the data architecture has no place to put that information.

In criminal justice, algorithms cannot account for caregiving responsibilities, experiences of gender-based violence, economic coercion, or the social contexts that shape women's trajectories through the justice system. Employment gaps caused by caregiving are scored as instability. Financial dependence resulting from economic abuse is scored as risk.

In both systems, the most consequential features of women's lives are invisible to the model — not because they are unmeasurable, but because no one built the infrastructure to measure them.

4. The Failure of Retrofitted Fairness

Our 2019 paper warned against treating bias as a technical problem with a technical fix. Both domains now confirm this warning.

In healthcare, AI systems trained on misrepresentative data reduced diagnostic accuracy by 11.3 percentage points compared with baseline clinical performance. Efforts to mitigate this through Explainable AI — providing explanations alongside biased outputs — produced the same 11.3-point drop. If the foundation is biased, explaining the bias does not correct it.

In criminal justice, removing gender as an explicit variable does not eliminate gender discrimination; it routes discrimination through proxy variables. Anonymization does not solve knowledge gaps. Post-hoc fairness patches create the appearance of equity while the underlying model continues to operate on male-default assumptions.

In both systems, fairness cannot be bolted on after the fact. It must be built into the evidence base, the data architecture, and the design process from inception.

IV. Where Bias Compounds: The Intersectional Architecture

The patterns described above do not affect all women equally. Algorithmic discrimination compounds along intersectional lines — race, ethnicity, indigeneity, disability, migration status, gender identity, socioeconomic position — and the evidence from both healthcare and justice shows that the women most harmed are those who were already least visible in the systems that generated the data.

In Criminal Justice

The compounding is measurable and severe. ProPublica's investigation of COMPAS found that Black defendants were 77% more likely to be rated high risk for violent recidivism even after controlling for age and gender. In the Netherlands, the HART system rated Moroccan and Turkish immigrant women 40% higher risk than Dutch women with identical criminal histories — bias stacking along gender, ethnicity, and migration status simultaneously. In Canada, Indigenous women were rated 35% higher risk despite significantly lower violent reoffending rates. In Australia, Aboriginal women were systematically over-scored through the same compounding mechanism. In the UK, transgender defendants faced 60% higher risk ratings than cisgender defendants with identical histories. In Germany, trans women were classified using male risk factors regardless of gender identity.

These are not isolated data points. They are the predictable result of training algorithms on data generated by systems in which policing, prosecution, and sentencing already operate along racial and gendered lines. The algorithm does not introduce new bias. It inherits every existing bias at once — and multiplies them.

In Healthcare

The same compounding operates through the clinical data pipeline. Joy Buolamwini's Gender Shades research revealed error rates of 34.7% for darker-skinned women in facial recognition systems compared with 0.8% for lighter-skinned men — establishing that intersectional accuracy gaps are the norm, not the exception, in AI systems trained on unbalanced data. In healthcare AI, chest X-ray models trained on datasets where men are consistently overrepresented learn correlations between disease and male anatomy, imaging context, and procedural patterns. Women with smaller heart size, different fat distribution, or breast tissue density are not simply underrepresented — their physiology is actively misread.

For women of colour, these gaps compound further. Diagnostic thresholds calibrated on white male physiology misperform for white women; they misperform more severely for Black women, Indigenous women, and women from communities where historical exclusion from research is deepest. Conditions that disproportionately

affect women of colour — lupus, sickle cell disease, fibroids — sit in the same data desert as endometriosis and PCOS, with fewer training examples, weaker labeling, and lower model accuracy.

The Structural Point

Intersectionality is not an additional consideration to be layered onto gender analysis. It is the analytical framework that reveals how algorithmic systems distribute harm. A model that is "fair on average" may be deeply unfair for Black women, Indigenous women, immigrant women, disabled women, transgender women — because aggregate performance metrics hide subgroup failures. This is hidden stratification: strong overall scores masking the fact that the system's worst errors fall on those who were already least served.

Both healthcare and justice systems must adopt intersectional evaluation as standard practice — not as a supplementary audit but as a design requirement. Training data must represent diverse identities. Performance metrics must be reported across compound categories. And the communities most affected must be present in design, testing, and oversight — because no statistical technique can substitute for the knowledge of people who live with the consequences.

V. What Must Change: An Integrated Agenda

The parallels between healthcare and justice reveal that the problem is not confined to either domain. It is structural, cross-cutting, and requires coordinated intervention across the full pipeline — from what we choose to study, to what we choose to measure, to who is at the table when systems are designed.

A. Rebuild the Evidence Base

The upstream problem is the same in both domains: the foundational evidence was generated without women.

In clinical research: Mandate sex-stratified analysis in all publicly funded trials. Require sex annotation for cell cultures and preclinical models. Expand post-market surveillance to disaggregate adverse events by sex. Invest in dedicated research on conditions that disproportionately affect women — endometriosis, autoimmune disorders, perimenopause — where data scarcity makes AI unreliable.

In criminal justice research: Fund longitudinal studies on women's trajectories through justice systems that capture caregiving context, economic coercion, and gender-based violence history. Develop sex-specific recidivism models rather than applying male-normed instruments to women. Ensure training datasets represent diverse gender identities, including non-binary and transgender individuals.

B. Redesign Data Infrastructure

Neither system currently captures the information needed to serve women equitably.

In healthcare: Build EHR modules for female health tracking — menstrual cycle characteristics, pregnancy history, menopausal status, hormonal therapies — as structured, time-varying fields. Elevate patient-reported outcomes to first-class data. Develop new data taxonomies that encode life-stage and cyclical dynamics rather than treating "female" as a static category.

In criminal justice: Develop risk assessment frameworks that account for caregiving responsibilities, experiences of gender-based violence, and the distinct pathways through which women enter the justice system. Ensure that employment, housing, and relationship variables are interpreted in context rather than scored against male-normed baselines.

C. Mandate Transparency and Accountability

Across both systems:

- Require sex-stratified performance reporting for all AI tools deployed in clinical care and judicial decision-making. Every model should disclose sensitivity, specificity, and error rates disaggregated by sex and, where possible, by

intersectional categories.

- Mandate independent algorithmic audits with public disclosure. Proprietary claims cannot override the right to know how automated systems reach decisions about health and liberty.

- Establish meaningful avenues for individuals to contest algorithmic determinations — the right to know when an automated system has contributed to a decision, how it reached that decision, and on what basis it can be challenged.

- Require ongoing monitoring for subgroup drift. Bias is not a static property; it can emerge or deepen as populations and data distributions change over time.

D. Enforce Regulatory Standards

In healthcare: The FDA's current guidance encouraging sex-disaggregated performance reporting is insufficient. Pre-market approval should require sex-stratified performance tables. Post-market surveillance should mandate sex-disaggregated adverse event reporting. The EU AI Act's classification of healthcare applications as high-risk must be matched with specific requirements for sex-aware validation.

In criminal justice: The EU AI Act's classification of criminal justice applications as high-risk provides a foundation but requires gender-specific implementation standards. National AI governance frameworks must incorporate human rights and intersectional lenses — not risk-only approaches. The licensing and relicensing of technology platforms should be leveraged as an accountability mechanism, as CEDAW Committee members have proposed.

E. Include Women in Design

Our 2019 paper argued that the exclusion of women from defining the rules of new systems would produce systems that exclude women. This remains true.

Both healthcare AI and judicial algorithms require diverse development teams that include gender studies scholars, community advocates, and representatives of affected populations — not consulted after key decisions are made, but present at the design stage. Gender-responsive design cannot be retrofitted.

The World Economic Forum found a 72% gender gap among AI professionals. Only 12% of researchers at leading machine learning conferences were women. At major technology firms, women comprise 10–15% of AI research staff. The systems shaping women's health and liberty are being built almost entirely without them.

F. Strengthen International Cooperation

Both domains are governed by a patchwork of national regulations and international norms that have not kept pace with deployment.

CEDAW's General Recommendation No. 33 on women's access to justice provides binding obligations when legal systems produce discriminatory outcomes — algorithmic or otherwise. The forthcoming GR41 on stereotyping will directly address how algorithmic systems reinvent and scale gender stereotypes. The UN Working Group on Discrimination Against Women and Girls will present its thematic report on AI and gender equality to the Human Rights Council in June 2026.

These mechanisms exist. What is missing is implementation, resourcing, and political will. Mass-scale correction of biased data systems requires multilateral cooperation. No region working alone can address what is a global pattern of exclusion encoded into global technology.

VI. The Stakes

In 2019, we wrote: *We are at a critical turning point. We can either seize this moment to correct bias in the digital realm, as we tackle bias in the analog world, or condemn ourselves to old bias hardwired into the future.*

The turning point has arrived. Healthcare AI is being deployed into clinical workflows that will shape diagnostic and treatment decisions for billions of people. Judicial algorithms are being embedded into court systems that determine liberty. In both, the evidence shows that women are being systematically underserved, misdiagnosed, disbelieved, and over-penalized — not because the technology is malicious, but because it learned from systems that were never built for them.

The technology is not destiny. These are human choices — what data to collect, what to prioritize, what to ignore. Better choices can change the course.

But the window is closing. Once biased systems are embedded as infrastructure, once their outputs become the training data for the next generation of models, the cost of correction rises and the likelihood of reform falls. We are building the permanent architecture of automated decision-making right now.

The question is whether women will be visible in it — or invisible by design.

This paper draws on three publications: "We Shape Our Tools, Thereafter Our Tools Shape Us: Artificial Intelligence, Automated Decision-Making & Gender" (Women At The Table, 2019); "Invisible by Design: Women's Health as the Blind Spot in AI and Medicine" (Women At The Table & FemTechnology, 2025); and "Gender Bias in Judicial Algorithms: A Global Analysis of Algorithmic Discrimination" (Women At The Table, 2026, CSW70 Expert Paper). It was prepared by Caitlin Kraft-Buchman, CEO/Founder, Women At The Table, and Oriana Kraft, Founder, FemTechnology.

About Women At The Table Women At The Table is a Geneva-based systems change, equality and democracy think tank focused on ensuring digital public infrastructure is inclusive by design. Their initiatives include the <A+> Alliance for Inclusive Algorithms and the <AI & Equality> Human Rights Initiative, partnering with

UN agencies, governments, and civil society globally to advance AI policy, applied research, and human rights-based frameworks.