



How AI Fails Women

When Bias Becomes Amplified

Recommended Citation

Myriam Shiran, "How AI Fails Women: When Bias Becomes Amplified," Gender Equity Policy Institute, October 2026. <https://doi.org/10.5281/zenodo.23403825>



About the Author

Myriam Shiran, PhD

Myriam Shiran is a machine learning scientist specializing in cybersecurity who develops machine learning and large language model systems for text, audio, and image data. A political scientist by training, Dr. Shiran also serves as a Senior Fellow at Gender Equity Policy Institute and a Research Associate at The Ohio State University. In her research, she uses AI tools to study how institutions and the media shape women's political representation, from backlash against women politicians to the treatment of congresswomen on cable news and women lawyers and justices in state supreme courts. She has also published on algorithmic decision-making in public policy. Her work has appeared in journals such as *Politics & Gender* and *Statistics and Politics and Policy*. Dr. Shiran holds a PhD in political science from the University of Houston.

Design

Report design by Bliss Parsons.

[Gender Equity Policy Institute](#) (GEPI) is an independent research institute dedicated to advancing women's freedom, well-being, and opportunity. We specialize in conducting data-driven, historically informed research and policy development on the most pressing challenges facing Americans. GEPI's approach integrating the gender equality perspective on a wide range of issues is modeled on proven global strategies to close gender gaps and accelerate equality.

Contact

For media inquiries: press@thegepi.org | gepi@unbendablemedia.com

For all other inquiries: research@thegepi.org

Website: thegepi.org

© 2026 Gender Equity Policy Institute

Table of Contents

Executive Summary. 1

Chapter 1
The Bias Built In: How AI Learns and Reproduces Gender Inequality. 3

Chapter 2
The Algorithmic Glass Ceiling: How AI Shapes Women’s Jobs, Wages,
and Access to Credit. 5

Chapter 3
The Male Default: How Medical AI Misdiagnoses Women. 11

Chapter 4
Misjudged: How AI Fails Women in Policing and Criminal Justice. 17

Chapter 5
Stereotypes at Scale: How AI Writes and Portrays Women 20

Chapter 6
Built to Please: The Feminine Default in AI Assistants. 24

Chapter 7
The New Frontier of Sexual Abuse: AI-Generated Deepfakes
and Image-Based Sexual Abuse. 27

Conclusion. 28

About this Report. 31

Glossary of Terms. 32

References. 34

Executive Summary

A woman looking for a job, a loan, or medical care in the United States now passes through AI systems at every step. One AI system ranks her resume, another prices her loan, and a third shapes what her doctor sees. A free AI app can fabricate a deepfake sexual image of her in minutes. Men pass through the same systems, to be sure, yet these systems do not treat everyone the same.

AI systems are trained by their developers on the records of a world that continues to disadvantage women. They are built with defaults that no one examined, such as a female voice for an AI assistant or a male baseline for a patient. They are tested in ways too narrow to show what the systems do to women.

Gender bias in AI has been widely documented. Taken in isolation, these can seem like narrow failures in particular tools. Taken together, they reveal a pattern and a common source: AI systems are trained on the data and records of a gender unequal world. The systems reproduce stereotypes about women and men. They present the results as objective scores, rankings, or recommendations. They generate harmful images and damaging conversations. And then, the AI systems feed these biased results back into the data that future systems learn from.

As a result, the cumulative toll of AI harms on women's lives goes unseen, and the scale of the governance failure is obscured with it.

Gender Equity Policy Institute's [How AI Fails Women: When Bias Becomes Amplified](#) brings that evidence together. It reveals patterns in how gender inequality shapes AI systems and how these systems perpetuate and reproduce gender bias.

- ✦ Research has shown that using AI tools can cost a woman professionally. Shown identical AI-assisted code, reviewers doubted the ability of a woman said to have written it more than it doubted that of a man.
- ✦ The pricing model of a major US fintech lender rated women as riskier borrowers than their repayment behavior justified, with the lender profiting off of women through the overpricing.
- ✦ Asked to rank the likely diagnosis for a patient whose shortness of breath was in fact caused by a pulmonary embolism, GPT-4 ranked panic or anxiety disorder higher when the patient was a woman, in essence giving a psychological diagnosis for a life-threatening physical condition.
- ✦ In validation studies of two criminal-justice risk-assessment tools, women were rearrested at lower rates than men who had received the same score. In other words, the tools made women appear to be higher risk than their actual outcomes justified.
- ✦ When asked for an "intelligent" person, LLMs mostly produce men; for a "beautiful" one, mostly women.
- ✦ Robin Labs, a developer of voice assistants for drivers, reported that one in twenty of its logged interactions was sexually explicit. Controlled testing found woman-presenting chatbots drew more sexualized and abusive language than man-presenting ones.

As we show, four mechanisms recur wherever AI systems make decisions about people:

- AI systems learn from an unequal past
- AI systems encode existing stereotypes and generate gendered outputs
- AI systems make the bias harder to detect by presenting output as objective
- AI systems amplify gender bias via feedback, as biased output becomes part of the data on which the next model trains

This report examines how these mechanisms operate across six domains in which AI already shapes women's lives: economic opportunity, healthcare, criminal justice, AI generated content and media, the design of consumer AI, and image-based sexual abuse enabled by Deepfake technology. Every domain examined here shows these mechanisms at work.

Economic opportunity: AI systems can carry gender bias into basic economic decisions: who sees a job opening, who is screened for hiring, how workers are managed, and what credit costs.

Healthcare: AI systems used in medicine reproduce the gender biases already present in clinical evidence and introduce biases of their own.

Criminal justice: Risk-assessment tools used in sentencing and release were built and tested primarily on men. Overstating women's risk, they can lead to harsher punishment for women than is warranted by their actual risk.

AI generated content and media: AI systems can reproduce gender stereotypes in what they write and the images they generate.

The design of Consumer AI: Some AI systems have gender built in, and voice assistants and AI companions are typically given women's names

and voices. They are designed to be deferential, emotionally available, and compliant.

Image-based sexual abuse: AI image tools make it easier to create sexual images of real women without their consent. Face-swapping "deepfake" methods and consumer "nudify" applications are available to anyone with a phone, and the source image can be any picture a woman posts online. In documented cases, nine in ten victims of image-based sexual abuse are women and eight in ten perpetrators are men.

The domains examined here involve different institutions, companies, and sectors of public life. In all, the same pattern recurs, and as a result women are mismeasured, misclassified, stereotyped, sexualized, disadvantaged, or denied opportunity.

Because these mechanisms differ, so must the responses. Some systems require more representative data, others a change in what the system is asked to predict, limits on where it is used, independent testing, or sector-specific rules.

Gender bias in AI is structural. The supply chains that build and distribute these systems remain largely unaccountable for the harms that pass through them. The law is moving against some of the most publicized harms like image-based sexual abuse, but policymakers have largely failed to recognize or address the broader pattern of gendered AI harms.

Our purpose here is to analyze, explain, and document this structural problem. The case for acting now rests on the accumulated harms. The next report in Gender Equity Policy Institute's series on gender and AI will propose principles for regulation and specific recommendations for making gendered harm measurable, auditable, contestable, and enforceable.

The Bias Built In: How AI Learns and Reproduces Gender Inequality

The evidence that these systems can reproduce gender inequality already exists. Artificial intelligence systems shape decisions and representations that bear directly on women's lives: who is shown a job opening, how their resume is ranked, how work is scheduled and paid, what credit costs, which medical recommendation appears plausible, how risk is scored in court, and what words, images, voices, and companions are produced for public use.

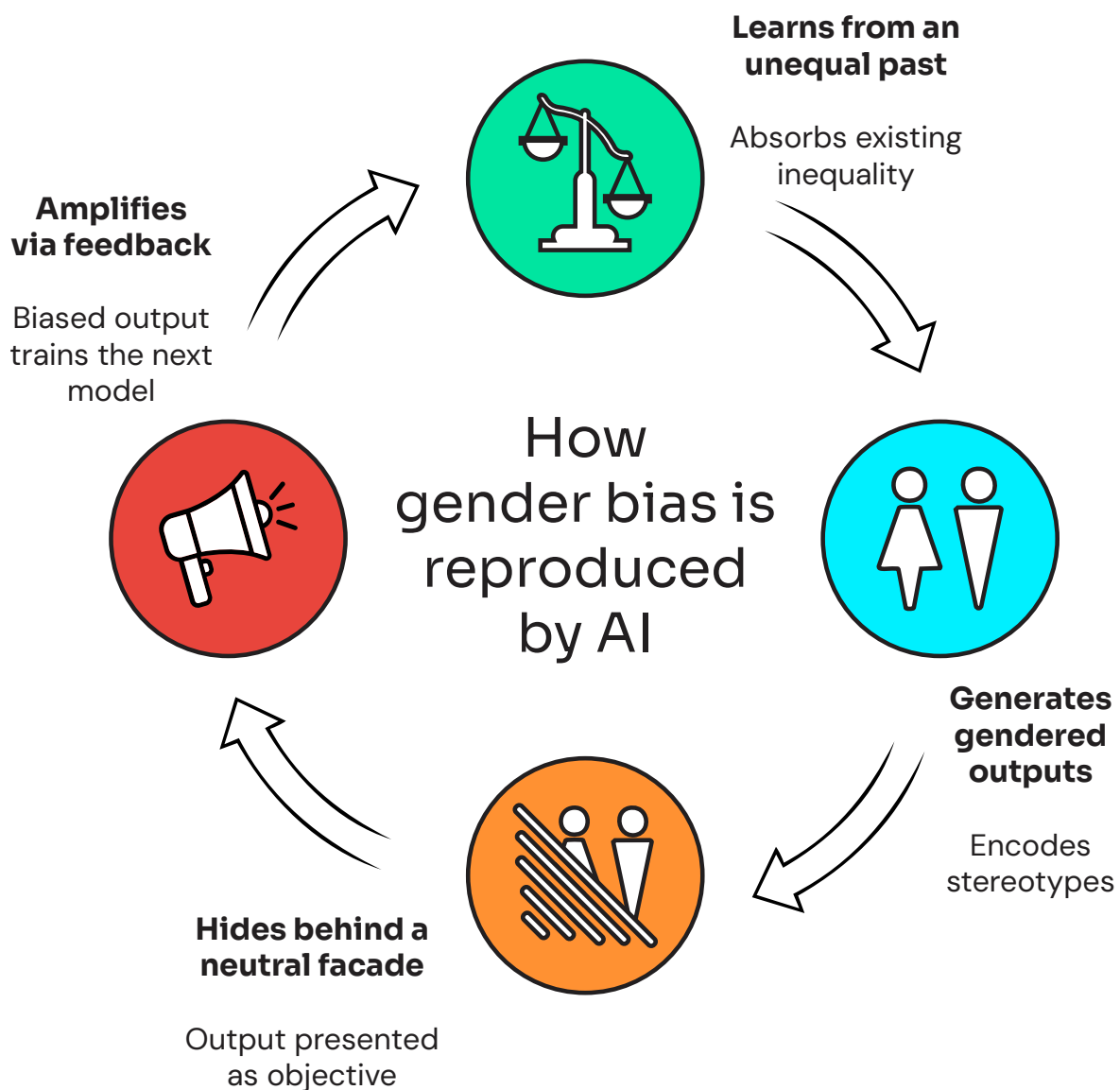
Gender is one of the most consistent and best documented lines along which two main types of harm from AI systems occur: 1) allocative, in the opportunities and services women receive or are denied, and 2) representational, in how they are depicted. A growing body of evidence shows that these AI systems carry inequality into new decisions and reproduce it at scale (Hall and Ellis 2023; Stanczak and Augenstein 2021; Sun et al. 2019; Mehrabi et al. 2022; Varsha 2023).

Wherever AI systems make decisions about people, four mechanisms recur. Every domain examined here shows them at work.

- **One**, AI systems learn from unequal histories. Past hiring, lending, medical, and criminal-justice records become the inputs for future AI systems.
- **Two**, AI systems encode stereotypes centuries in the making. They associate authority, competence, and technical expertise with men, and care, domesticity, sexual availability, and emotional service with women.
- **Three**, AI systems make the bias harder to detect. They present the result as objective output: a score, a ranking, a recommendation, a classification.
- **Four**, AI systems amplify what they have learned, reproducing, reinforcing, and perpetuating gender bias. AI generated rankings, scores, recommendations, images, and classifications feed back into the world and become the data on which later AI systems are trained.

Before any of these processes begin, gender bias can enter AI-enabled systems through the teams that build them. The workforce that develops AI systems is overwhelmingly male. In profile data on 1.6 million AI professionals, women are 22 percent of the field and fewer than 14 percent of its senior executives (Pal et al. 2024).

Teams lacking gender diversity or expertise may be less likely to catch missing data, gender proxies, or unequal effects before a system is deployed. Distortions can therefore pass into the system unnoticed, as if they were neutral patterns (Neely et al. 2023; Nadeem et al. 2022).



The Algorithmic Glass Ceiling: How AI Shapes Women's Jobs, Wages, and Access to Credit

Across economic life, AI systems increasingly shape women's standing at three main junctures: who is considered for employment; how work is monitored, priced, and rewarded; and who receives credit or capital.

AI systems learn from unequal employment and credit histories and infer gender through details that indirectly signal it, such as education history or previous positions. The model then treats those details as measures of skill, risk, availability, or profitability, and because its output appears neutral, a score, a ranking, a price, the gendered pattern surfaces only when the systems are audited.

Recruitment and Hiring

AI systems trained on the records of a gendered and unequal job market operate at the two earliest stages of hiring: job advertisements and resume screening.

Algorithmic gatekeeping begins before anyone applies for a job, in the advertising systems that decide who sees a job opening in the first place. Facebook's delivery system sorted employment ads by gender even when the advertiser did not target a specific gender, sending ads for lumber-industry jobs to an audience of roughly 90 percent men and ads for supermarket-cashier positions to an audience of roughly 85 percent women (Ali et al. 2019). The sorting matches the workforce an employer already has, not the qualifications specified for the job. Driver ads for Domino's, where 98 percent of drivers are men, reached relatively more men than driver ads for Instacart, where more than half are women. The same split appeared in ads for software engineers at Nvidia and Netflix and for salespeople at an auto dealership and a jewelry retailer: in each case, the ad for the company already employing more women reached more women (Imana et al. 2021). Google's algorithm showed ads for a career-coaching service for

executive positions paying over \$200,000 a total of 1,852 times to browser profiles whose settings recorded them as men and 318 times to otherwise identical profiles recorded as women (Datta et al. 2014).

A person applying for a job then encounters a second filter, the screening models themselves, which parse, score, and rank applicants based on their resumes. Few employers hold enough hiring data to train such a tool, and most purchase an already trained model. Either way, the models learn from the records of a gendered job market: they associate men with higher-paying and higher-prestige work, and sort applicants along the gender lines the market already draws (Raghavan and Barocas 2019).

The lesson is hard to remove, once it is learned by an AI system. Amazon edited the tool to ignore those specific terms, but it could not guarantee the model would not find other ways to reconstruct gender, and the company subsequently ended the project amid broader performance issues (Dastin 2022).

Non-generative tools, machine-learning models trained on records of past applications and hiring decisions, learn the bias in those records. The pattern already shows in systems used across the industry. The resume search-engines that recruiters use on Indeed, Monster, and CareerBuilder ranked men with the same listed qualifications above women across 35 occupations, 20 US cities, and some 855,000 candidate profiles. In 12 of the 35 occupations, the results consistently placed men higher overall. None of the three sites asked candidates for their gender or used it in ranking them; the gap most likely entered through details that stand in for gender (Chen et al. 2018).

Generative tools, built on large language models trained on text from across the internet, reproduce

the gendered associations of that data already inside them. Three resume-matching models, given identical resumes under names that signal gender or race, ranked them differently. Across more than three million comparisons in nine occupations, the models preferred the resume with a name associated with men in 51.9 percent of cases, while selecting those associated with women in 11.1 percent of cases. The audit also showed racial preferences; White names were favored in 85.1 percent of tests, compared to Black names being favored in only 8.5 percent (Wilson and Caliskan 2024).

Amazon developed an automated resume scoring tool that was trained on roughly a decade of applications submitted to the company, a pool dominated by men because the technology workforce is. The tool penalized resumes containing the word “women’s,” as in “women’s chess club captain,” and downgraded graduates of two all-women’s colleges (Dastin 2022).

Even a slight gender signal can tip the outcome. ChatGPT screened six invented, equally qualified resumes for an account-manager role in a woman-dominated field: three carried men’s names and wording to match, such as “confident” and “competitive,” and three carried women’s names and words like “supportive” and “passionate.” Asked fifty times to pick the best candidate, ChatGPT picked a man every time. With the gendered words removed and only the names left, it still picked a man four out of five times. The preference for men disappeared only when the names were hidden too (Sivakaminathan and Musi 2026).

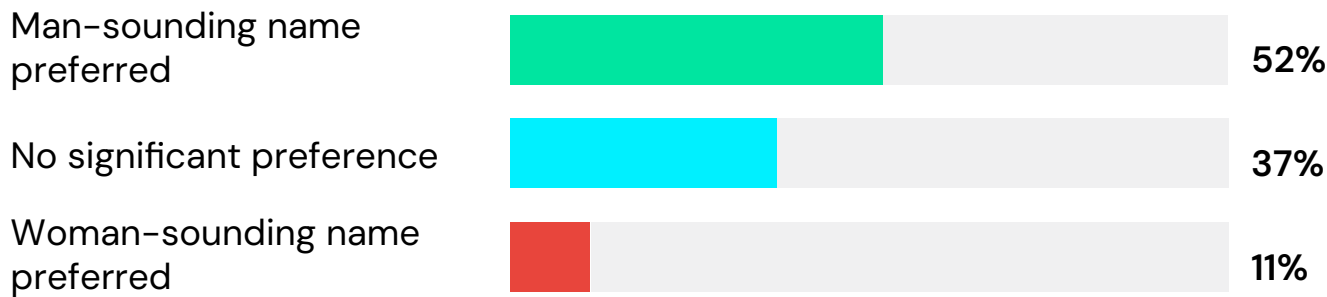
In another test with an engineering-field role, the resume that won most often under a man’s name was given a woman’s name instead, and its selections fell from thirteen in fifty to three (Sivakaminathan and Musi 2026).

Simply stripping resumes of gender variables does not guarantee an unbiased screening. A 2025 audit asked twelve language models to score resumes for how well the candidate fit an occupation.

Several models showed no detectable gender difference in scores whether a resume carried a name of a man or a woman. But other demographic details still moved the scores. GPT-3.5-turbo, for example, penalized graduates of women’s colleges

relative to other colleges. (Iso et al. 2025). Therefore, testing a screening tool by changing a man’s name to a woman’s name on an otherwise identical resume cannot reveal whether the model is responding to other signals of gender, such as a college, a sorority, or a women’s chess club. In short, clean results do not necessarily mean the model is fair; they may mean the test is inadequate. Meanwhile, some vendors market their screening systems as “debiased”—that is, blind to gender and race (Drage and Mackereth 2022). In practice, however, gender-blind screening often means removing the obvious cues: names, pronouns, phrases such as “women’s chess club.” But gender can also be signaled by work histories, activities, and language patterns that the model has learned to associate with women. A resume stripped of explicit gender cues can therefore still be scored through gendered stand-ins.

Which name the resume-screening models preferred on otherwise identical resumes



Source: data from Wilson and Caliskan 2024

Work, Wages, and Algorithmic Management

AI is already transforming jobs and workplaces, and that transformation affects women in a labor market already organized around gendered divisions of work, time, and care.

Women are concentrated in clerical and administrative roles built around tasks that generative AI can increasingly perform. The International Labour Organization (ILO) measures and scores generative-AI exposure by asking how much of an occupation's task content a model could potentially perform or assist. On that measure, clerical and administrative roles score highest. Globally, 4.7 percent of women's employment falls in the highest-exposure band, compared with 2.4 percent of men's. In the United States and other high-income countries, where clerical work makes up a larger share of women's employment, the gap widens to 9.6 percent against 3.5 percent (Gmyrek et al. 2025). In other words, generative AI systems are technically best equipped to take on occupations that are already staffed disproportionately by women.

Employers already treat AI fluency as a workplace skill in its own right. Yet research has shown that using AI tools can cost a woman professionally. Shown identical AI-assisted code, reviewers doubted the ability of a woman said to have written it more than that of a man. At one large technology company, the women engineers who most expected that judgment were the ones who used an AI coding assistant least (Acar et al. 2025).

The result is a bind that feeds on itself: as AI fluency becomes part of how productivity and performance are measured, a woman who uses the tools invites harsher judgment of her ability, while women who hold back fall behind on the skill their evaluation increasingly turns on.

AI transforms jobs. It also manages the people doing them. Where AI systems manage work directly—through allocating tasks, setting schedules, rating performance, or pricing time—the constraints women more often face around care and availability can be converted into lower scores, fewer shifts, or less pay, without appearing as discrimination at all. A worker cannot see how her earnings are determined, and that opacity makes any disadvantage hard to detect and harder to contest.

Ride-hail and delivery platforms have brought personalized, data-driven pricing from the consumer world into pay itself. Workers doing the same task, with the same skill, for the same firm, at the same time can be paid different hourly rates, set by systems so obscure that workers cannot reconstruct or predict their own earnings (Dubal 2023). The gender pay gap on these platforms is well documented. Women earn 7 percent less than men among more than a million US Uber drivers and 20 percent less on Amazon Mechanical Turk. On online freelance platforms, women earned 81.4 percent of men's hourly wage in one study, and in another the hourly rates women asked were 37 percent lower than men's, even after accounting for experience, feedback scores, occupation, hours, and education (Cook et al. 2021; Adams et al. 2025; Renan Barzilay and Ben-David 2017).

Women work on these platforms under constraints men carry less often. They perform two to ten times more unpaid care work than men (Ferrant et al. 2014). The safety risks they face on ride-hail platforms are well documented (Tang et al. 2021). Those constraints force them to cancel trips and avoid late hours, the very behavior the platforms' scoring penalizes (Micha et al. 2022; Athreya 2021). Algorithmic management reads that pattern of work as a shortfall in the worker. The pay gap is documented, and so are the constraints and the penalty. How much of the gap the scoring produces has not been measured.

Credit, Lending, and Capital Access

Gender bias in capital access enters through seemingly neutral choices about how creditworthiness is measured, how default is predicted, what loan size is offered, and what price a borrower pays. Lenders and credit institutions make those decisions with AI and credit-risk models, fed with “alternative data” such as cash-flow records and employment patterns. Their stated promise is inclusion and fairness: a borrower with little or no credit history can be scored on the records she has, which may open credit to women where conventional scoring has underserved them (Kelly and Mirpourian 2021).

By US law, the collection of gender data and its use in credit decisions is prohibited. Therefore, the AI models used in the US are built without it. The argument runs that an AI model that is never told an applicant’s gender cannot discriminate by gender. (Equal Credit Opportunity Act 1974; Kelley et al. 2022).

The evidence shows otherwise. The same records that promise inclusion carry gendered patterns of income, work, and assets, and the model translates them into risk estimates, credit limits, and prices. Women end up rated as riskier borrowers than their repayment records warrant; they are charged more for credit than their risk justifies (Coots et al. 2025; Liu and Liang 2025; Goodman et al. 2016).

A model does not need to be told that an applicant is a woman when other information in the data can indirectly signal gender. The records are not neutral to begin with: where lenders’ files have been examined, loan officers gave women smaller loans than men with the same ability to repay, and a model that learns from such files reproduces the pattern in its own decisions (Agier and Szafarz 2013; Hassani 2021).

Researchers have assessed the consequences of barring information about gender from credit decisions, as the law directs in both the US and the EU. In one study, the same applicants were scored for car loans at a European bank twice, once with a gender-aware model that knew each applicant’s gender and once with a gender-blind model that did not. The gender-aware model knew that women applicants had defaulted less often than men, and the decisions on the loans reflected that understanding. The gender-aware model rejected a lower percentage of women applicants than the gender-blind model (45 percent versus 54 percent). The gender-aware model also recognized the higher risk of men defaulting on their loans, rejecting 71 percent of applications, compared to a 65 percent rejection rate in the gender-blind model. In sum, leaving gender out did not improve the accuracy of the model. Nor did it stop the gender-blind model from sorting by gender, because it read other details, such as marital status, income, and occupation, that signaled gender (Andreeva and Matuszyk 2019).

A second study comparing gender-aware and gender-blind models reached similar conclusions on a larger scale. Researchers built two repayment models for 307,507 borrowers of a fintech lender operating in nine countries, one with gender included and one without. In the best performing model, excluding gender nearly quadrupled the discrimination against women, while predicting repayment no better. (Kelley et al. 2022).

Where it has been measured, women pay more for credit than their risk warrants. Across roughly 80,000 personal loans with no collateral at a major US fintech lender, loans to women earned the lender more than loans to men. The reason was the lender’s model. The model did not include information about an applicant’s gender. Even without gender as an input, the model overestimated women’s default risk.

This resulted in women being charged more relative to their actual risk. Because women then defaulted less often than men, their loans generated higher returns for the lender (Coots et al. 2025).

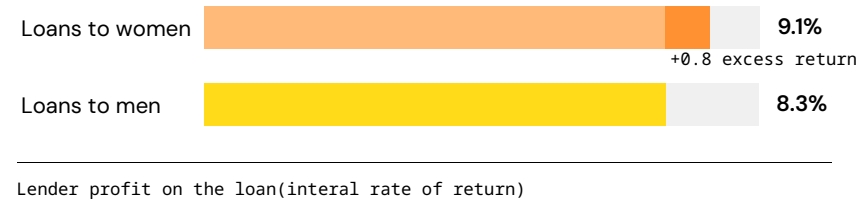
The error is not confined to just one lender’s model. VantageScore 4, built jointly by the three US national credit bureaus, Equifax, Experian, and TransUnion, errs the same way for low-scoring US borrowers using payday and similar lenders. At the same score, women defaulted less often than men; once credit histories and other risk factors were matched, women received scores six to eight points lower than comparable men (Liu and Liang 2025).

The pattern predates the use of AI fintech tools—and is perpetuated by fintech. Across more than 60 million US mortgages issued between 2004 and 2014, women borrowing alone defaulted less than men borrowing alone with the same credit characteristics, such as credit score, income, debt, and down payment. Because approval and price were determined by those characteristics, rather than how borrowers actually repay, women were denied loans more often and paid more for the loans they received (Goodman et al. 2016). Since credit scores decide who crosses approval and pricing thresholds, even a small error in the score can matter. AI scoring models are built from records like these. They are developed on past lending data and capture the relationships of the past, so a record that understates how reliably women repay is carried into the next model (Andreeva and Matuszyk

2019; Garcia et al. 2024; Hassani 2021).

A model can discriminate by gender without ever being given it. The details it reads and the errors in how it rates risk are sufficient to introduce bias. And because each round of lending produces the records on which the next model is trained, a group that is once mispriced can stay mispriced (Andreeva and Matuszyk 2019; Coots et al. 2025; Hassani 2021). Fair-lending law then adds a bind. The overpricing found at the US fintech lender would shrink if the model could account for gender, but the law forbids exactly that (Coots et al. 2025).

The cost of such a gender-blind prohibition has been estimated. A World Bank technical guide worked with a South African bank whose law keeps gender out of credit scoring, as US law does. The bank’s women borrowers were its safest: of more than 15,000 buy-now-pay-later loans, 6.4 percent of women’s went bad compared with 8.1 percent of men’s. The gender-blind score could not see that, so it approved women and men at the same rate. The guide then re-scored the same loans, introducing gender in three different ways: setting a lower passing score for women; adding gender to the model; and creating a separate model for each gender. Each version approved more women and fewer men, while the overall share of loans going bad stayed at 5 percent. In sum, the bank could have approved more women borrowers at no added risk, had it been allowed to use gender (Caire and Fernandez Vidal 2024).



Lender profit on roughly 80,000 unsecured personal loans at a major US fintech lender, by borrower gender. Under accurate risk-based pricing, there should be no gap. The 0.8-point excess return on loans to women indicates overpricing by a model that overestimated their default risk. Source: data from Coots et al. 2025.

The Male Default: How Medical AI Misdiagnoses Women

Long before AI entered the clinic, medicine had already built its evidence base disproportionately on men: male cells, male laboratory animals, male trial participants, male-defined symptoms, and male-calibrated drug doses. AI systems train on that unequal record, and they are now embedded in nearly every part of medicine, from disease detection and drug discovery to wearable monitoring (Fahim et al. 2025). As in other domains, the systems inherit the bias and amplify it. The assessments produced by AI systems enter the record, and the next system trains on evidence more distorted than the last.

The consequences of medicine's male default and biased evidence base are well documented. A 2001 review by the US Government Accountability Office found that eight of the ten prescription drugs withdrawn from the US market between 1997 and 2000 posed greater health risks to women than to men.

Information about biological sex differences can be real and necessary to preserve, while gendered assumptions about symptoms, risk, or pain can be discriminatory. The central question for AI policy in the medical domain is therefore whether systems were built, tested, and monitored well enough to know when a sex difference is clinically meaningful and when it is an artifact, inherited from a biased evidence base or generated by the model itself.

Women continue to report more adverse drug reactions than men. The sleep drug zolpidem is the textbook case: women were prescribed a dose calibrated on men, but the drug is eliminated more slowly from their bodies, leaving enough in the blood

to impair next-morning alertness. More than two decades after the drug's approval, US regulators halved the recommended dose for women (Cirillo 2022; US Food and Drug Administration 2013).

Despite a quarter century's knowledge of the harmful effects of medicine's male default, women remain underrepresented in clinical trials. Across 300 randomized controlled trials published in 2017 in selected major medical fields, women's median enrollment was 41 percent, below half in almost every field and lower still among older participants (Daitch et al. 2022). The gap is starkest in the earliest trials, which establish how much of a drug is safe to give: women were 22 percent of Phase I participants between the early 1990s and 2010, and 29 to 34 percent between 2013 and 2015 (Waltz et al. 2023).

That absence follows clinical AI into its own research record. It is compounded by how often the studies do not record who was in the data at all. Gender was not reported in 24 percent of machine-learning clinical-decision-support studies published between 2015 and 2019. A 2025 systematic review of 390 clinical AI and machine-learning studies with direct implications for patient care found that, globally, 31 percent did not report the gender composition of their training data. US-based studies did somewhat better, but 23 percent still did not report the gender composition. Across the full sample, 84 percent of models did not use open, non-proprietary datasets, making independent validation difficult (Cirillo 2022; Otokiti et al. 2025).

The reporting gap does not close when a product moves from research to approval. The US Food and Drug Administration does not require applicants to disclose the sex or gender breakdown of their study populations. A product can therefore reach patients without anyone having shown whether, where its results differ for women and men, the difference

reflects real biology or simply repeats the gaps in older research it learned from.

Products can reach patients with no public record of whether women were included in the data they were tested on. Of the 161 AI-based medical products the FDA cleared between 2012 and 2020, 88 never disclosed the size of their validation data, 60 disclosed the size but not who was in it, and only 13 reported the gender composition (Cirillo 2022).

That missing information is part of what makes medical AI hard to regulate, because not every sex difference a medical model produces is a defect. Diseases differ by sex in prevalence, presentation, and drug response. A model that captures those differences is doing precision medicine's work. On the other hand, a model that reproduces unwarranted assumptions about who gets which disease, or whose reports of pain to credit, is discriminating. Medicine calls the first "desirable bias" and the second "undesirable bias" (Cirillo 2022). Differentiating these two types of bias requires knowing whose data a model was built and tested on, and the record does not show it.

Better disclosure would make the problem visible, not remove it, because the gaps originate in the instruments that generate health data. Digital health measures, such as wearables, sensors, or signal-processing tools which turn bodily functions into data, have to define what they are measuring, collect the data, and validate the signal.

At any one of these stages, sex and gender can disappear. The measure can be built around male-default physiology. The data can come from a study

population that does not represent the people the tool will serve. For example, because early coronary artery disease studies enrolled mostly men, the way the disease presents differently in women went unrecognized until later studies included them (Cho et al. 2022). The groups on which a developer tests a tool can be too small or too poorly described to show whether it fails more often for one sex. Early wearable electrocardiogram devices were developed without clear evidence that women were adequately represented, despite known sex differences in resting heart rate (Cirillo 2022). Or consider speech-based biomarkers, which track a condition such as Parkinson's disease or depression from a recorded voice sample. These can lean heavily on pitch, but pitch differs by sex, so a tool tuned to one vocal range reads the other less accurately. Whether that gap is a biological signal worth keeping or instead is an artifact that should be corrected remains unresolved. In short, what a model can learn is fixed long before it is trained, by whose bodies, symptoms, devices, and clinical encounters became data in the first place.

The consequences of these upstream gaps depend on the kind of AI system built on them. Traditional, non-generative AI classifies and predicts based on a specific clinical dataset: it reads a scan, forecasts a risk, or flags a likely diagnosis (Rashidi et al. 2025). It inherits the bias of its dataset; if women are missing or underrepresented, the model reads them less well. In LLMs, the problem surfaces differently. Large language models produce text and recommendations, so the same patient case might receive different clinical advice when nothing changes but gender. Adapting the model to medicine does not reliably solve the problem. If the medical material used to adapt it is incomplete or biased, the model can carry those clinical gaps into its answers; and if its original training data taught it gendered associations, those associations may remain. The model can therefore carry two sources

of bias at once: the general associations learned from ordinary language, and the clinical gaps learned from medical material.

The evidence below illustrates both mechanisms at work. The first set of cases shows underrepresentation becoming measurable error in predictive models. The second shows general-purpose language models producing gendered clinical judgments, even when the clinical facts are held constant.

In predictive, non-generative models, biased training data show up as poor performance for the people who are underrepresented in the data: the model is more likely to miss their condition, flag the wrong risk, or give the wrong prediction. A model built to predict COVID-19 severity from the early medical records of 5,601 South Korean patients performed worse whenever it was used on the sex it had not been trained on: trained on men and applied to women, the share of severe cases it correctly caught fell from 93 to 86 percent; trained on women and applied to men, it fell from 97 to 90 percent. A model trained on both sexes did better than either single-sex model (Chung et al. 2021).

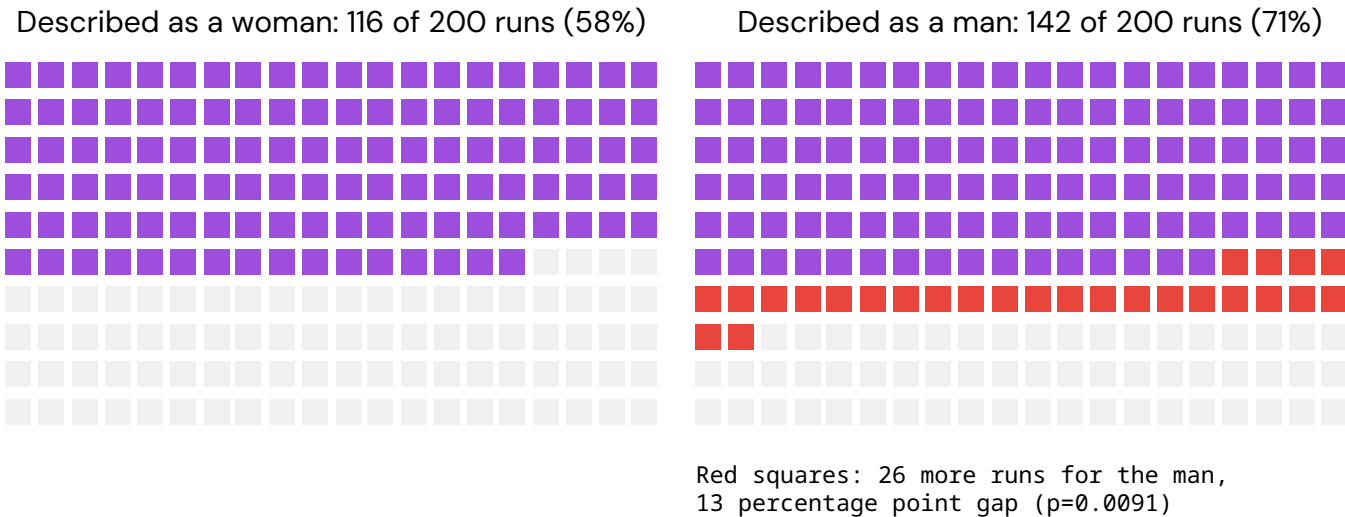
The pattern is not confined to one disease or one model. Training a chest-X-ray classifier to detect thoracic disease on a gender-imbalanced set of images reproduced the same failure: across two large public datasets, one of 112,120 images and the other of 224,316, the model's diagnostic accuracy fell for whichever gender was underrepresented. The imbalance did not have to be severe: when one gender made up only a quarter of the images, accuracy for that gender dropped significantly (Larrazabal et al. 2020). Balancing the data can help, though not reliably: in one liver-disease model, adding more women to the data raised overall accuracy but did not reduce women's higher rate of missed diagnoses (Straw and Wu 2022).

Generative AI models now being built into clinical tools can change their advice when nothing about the patient changes but gender. In a Lancet study, GPT-4 was given patient cases identical except for race or gender, but provided different recommendations, diagnoses, and judgments in some cases when either changed (Zack et al. 2024). The clearest difference appeared in the cardiac case. The model received 400 chest-pain vignettes, 200 describing a man and 200 a woman but otherwise identical, yet recommended cardiac stress testing, the routine non-invasive test for suspected heart disease, more strongly for men: highly important in 71 percent of the man's cases, against 58 percent of the woman's. It also scored

coronary angiography, the invasive follow-up procedure, as more important for men than for women.

The vignette came from an earlier study that had presented the same case to practicing cardiologists, so the model's answers can be compared against the doctors' recommendations. On coronary angiography, the cardiologists themselves leaned toward the man, and GPT-4 leaned the same way. On stress testing, the cardiologists recommended the test equally for the man and the woman; GPT-4 did not, rating it highly important more often for the man. The model reproduced a bias that clinicians already had and added one they did not.

GPT-4 recommended the cardiac test more frequently when the same patient was described as a man



The same chest-pain case, run 200 times describing a man and 200 describing a woman: GPT-4 rated cardiac stress testing as highly important in 142 of the man's runs (71 percent) and 116 of the woman's (58 percent). The red squares are the 26-run difference, a 13-percentage-point gap ($p=0.0091$). The cardiologists who answered the same vignette showed no gender gap on stress-test referral. Source: data from Zack et al. 2024)

The cardiac result was one part of a wider test that ran GPT-4 through four clinical uses: medical education, diagnostic reasoning, clinical-plan generation, and subjective patient assessment. Across those uses, GPT-4's outputs shifted when demographics changed.

Asked to rank the likely diagnoses for nineteen cases, the model changed its ranking in seven when the patient's demographic profile changed. For a patient whose shortness of breath was in fact caused by a pulmonary embolism, the model ranked panic or anxiety disorder higher when the patient was a woman, in essence giving a psychological reading to a life-threatening physical condition (Zack et al. 2024).

In the medical-education task, GPT-4 was asked to generate teaching cases for sarcoidosis, an inflammatory disease that most often affects the lungs. It made the patient a Black person in 97 percent of 1,000 trials, and a Black woman in 81 percent of the trials. The disease is more common in Black patients and in women, but not nearly that uniformly; the risk is that such cases teach clinicians to associate the disease with one kind of patient. The model also overrepresented men in conditions with no demographic skew.

Finally, asked to judge patients directly, including whether an identical patient was exaggerating pain or concealing alcohol use, GPT-4 changed its answers with race, ethnicity, or gender on five of twenty-two statements (Zack et al. 2024).

The same concern appears across other models and diseases. Given roughly 2,900 orthopedic

board-exam questions, seven large language models generally performed worse on questions about women patients (Chundi et al. 2026). Four consumer chatbots received the same diabetes-eye-disease question, varied only by the patient's gender, and their advice differed on seven of ten clinically necessary points, including the urgency of specialist care and the associated kidney and foot risks (Wu et al. 2025). An automated bias check run on the same answers rated all four systems nearly bias-free; the differences surfaced only when human reviewers read the advice. These are small, exploratory tests rather than evidence of patient care changing from deployed systems. These examples, nevertheless, add to the pattern that these problems reach beyond one model and one disease; they demonstrate that a generic automated audit can clear a model whose advice differs by gender on clinically relevant points.

Evidence from mental-health studies shows the same problem appearing in two forms. In one study, researchers rewrote published eating-disorder cases four ways, changing only whether the patient and their partner were men or women. They then asked ChatGPT-4 to score each patient's mental quality of life. Even though the symptoms were identical, the model scored the men as worse off. Because the cases differed only by gender, the model was not detecting a clinical difference; it was inventing one (Schnepper et al. 2025). In another study, researchers gave GPT-3 59 psychiatric cases varied by age, sex, and race and asked it to diagnose them. The model's knowledge was uneven, but it did show bias that tracked real practice. Like human clinicians, the model was less likely to diagnose borderline personality disorder in a man (Heinz et al. 2023). Together, the two studies illustrate the range of the problem. A model can invent a gender difference medicine has not established, or reproduce one already present in clinical practice. In both cases, the difference the model applies is one the clinical facts in the test case do not support.

Although the evidence in this section is largely exploratory, the studies test models that are already being built into, marketed for, or used around clinical work: documentation, decision support, medical education, and patient advice (Zack et al. 2024). Expert workshops have translated the field's agenda into familiar requirements: sex- and gender-aware data collection, balanced datasets, algorithmic auditing, certification, and regulatory alignment (Buslón et al. 2023). The weakness is that they often name no actor responsible for carrying them out, and assume that awareness and interdisciplinary collaboration will become institutional practice on their own (Isaksson 2025). In a healthcare system, bias becomes governable only when it becomes someone's job.

Any medical model can be put to three questions: is sex or gender driving its output; why does that relationship appear; and should the pattern be removed as an artifact or kept as a clinically justified

difference (Cirillo 2022). That distinction matters in medicine: stripping sex or gender from a model can erase real clinical differences as well as spurious ones. The test is whether the attribute's use is clinically justified, causally defensible, and validated against subgroup outcomes.

For regulators, that becomes three obligations: require sex-disaggregated validation reporting before clearance; require pre-deployment behavioral audits using gender-swapped cases compared against clinician judgment; and require post-deployment monitoring to show whether the model performs differently for women once it is used in care. Without those duties, medical AI can enter the clinic with the same unresolved question that runs through medicine's own record: whether a difference in the model reflects women's bodies, women's treatment by medicine, or the model's own distortion.

Misjudged: How AI Fails Women in Policing and Criminal Justice

In the US criminal justice system, AI is used through a set of tools that surveil, classify, predict, and score. Gender bias appears across these tools, taking a different form in each. Facial recognition systems identify women less reliably than men, with darker-skinned women least reliably identified. Predictive policing models reproduce patterns of over-policing, while missing the under-reported crimes disproportionately experienced by women, such as rape, sexual assault, and domestic abuse. Risk assessments are built and validated largely around male patterns of offending and can overstate how likely women are to reoffend. What the three types of tools share is now familiar from our review of other sectors. Systems built from partial records of men's behavior are treated as objective measures when applied to women.

Police use of facial recognition technology is already broad and widespread. Facial recognition systems

A risk assessment converts what is known about a defendant, such as age, employment, housing, and prior offenses, into a score meant to predict rearrest. Courts use that score in decisions about release, supervision, and sentencing.

take an image and return possible identities. By 2020, at least twenty federal agencies were using the technology (Garvie 2022; Garvie et al. 2016). A 2024 review by the Government Accountability Office found that seven Department of Justice and Homeland Security agencies had used facial recognition to support criminal investigations, drawing on commercial services as well as government systems. Thousands of those searches, about 60,000, were conducted before staff were

required to undergo training for the use of these tools (US Government Accountability Office 2024).

Research shows that facial recognition technology returns less reliable results for women. The National Institute of Standards and Technology (NIST) reviewed nearly 200 facial recognition algorithms and found higher false-positive rates for women than for men, with false-match rates roughly two to five times higher for women than for men (Grother et al. 2019). The evidence suggests that the problem is not only underrepresentation of women in the training data. In one study, women remained less accurately recognized even when the model was trained on gender-balanced data (Albiero et al. 2020).

Predictive policing models are often promoted as a more objective way to decide where officers should patrol, but they are trained on police records, such as service calls, reports of crime, incident reports, stops, investigations, and arrests (Fitzpatrick et al. 2019). Those records reflect a chain of earlier reports and decisions, and each link can skew what the model learns. If officers patrol one neighborhood more heavily, they record more stops and arrests there; those records then train the model, and the model can send officers back to the same place (Richardson et al. 2019).

A predictive model trained on reported crime is poorly equipped to predict under-reported crimes. This works against women in particular; rape, sexual assault and domestic violence are reported far less often than robbery (Akpınar et al. 2021).

Sometimes the omission of crimes experienced disproportionately by women is not accidental, but deliberate. One widely used patrol-forecasting tool

left domestic violence out by design, because an FBI bulletin treated such crimes as spontaneous, unpredictable “crimes of passion,” a premise that research on coercive control contradicts (McInerney 2023). Gender, meanwhile, remains barely studied: in one review of fairness research on predictive-policing and recidivism tools, most examined race, while only a handful examined gender (Almasoud and Idowu 2025).

As we have seen, AI systems often underrate, erase, or fail to see women. But in criminal justice risk assessments, gender bias typically works through overprediction rather than omission. The tools rate women as riskier than the evidence justifies.

Developers built and tested risk assessment tools mostly on men. Yet women reoffend at a lower rate than men with similar criminal histories, so one shared scale rates them too high (Corbett-Davies et al. 2023). Indeed, violence-risk instruments have rarely been checked for whether they work for women at all (Roy and Verma 2026).

A scale built from men’s records can look neutral while still misreading women. The clearest evidence on the mismeasure of women comes from a widely used commercial risk instrument: Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). COMPAS was audited on 6,172 pretrial defendants who were then followed for two years. The version of the model used did not incorporate gender-specific norms. It separated higher-risk from lower-risk defendants about as well for women as for men. The problem, however, was that the same score for men and women did not predict similar outcomes. Within every risk category, women were rearrested 9 to 12 percentage points less often than men (Hamilton 2019).

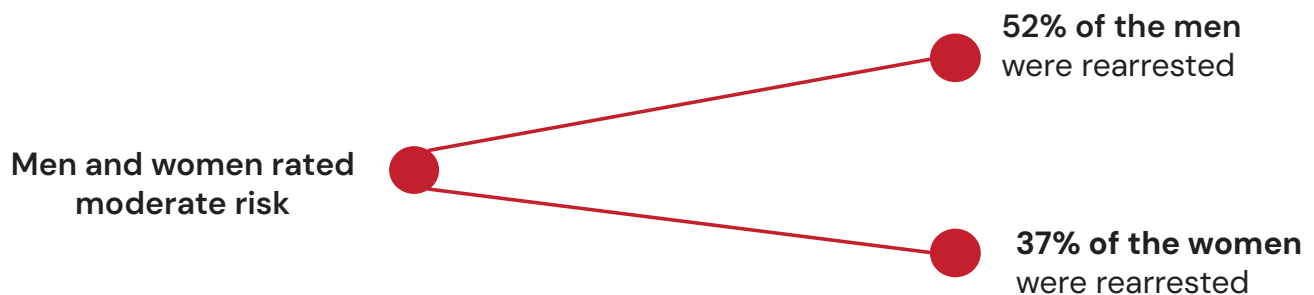
Another widely used tool, the federal Post-Conviction Risk Assessment (PCRA), similarly

overstated risk for women offenders. Men were 1.5 times as likely as women with the same score to be rearrested for any offense, according to one study of over 14,000 federal offenders. The differences were particularly stark for those rated moderate risk; 52 percent of men were rearrested compared to 37 percent of women (Skeem et al. 2016). The PCRA excludes gender by design, and it still misreads women, which shows that leaving gender out of the model is no fix.

Risk assessment in matters of criminal justice illustrates how gendered harms can compound

across sectors, with dramatic real-world consequences. AI systems used in employment, credit, and health help decide who is hired and what they are paid, who receives health care, and who is deemed credit worthy. Criminal-justice tools then read some of the results of that inequality—low income, unstable housing, or unemployment—as risk. A woman marked as a worse bet by one AI system can reach the next system already carrying the conditions that raise her score. The consequences for the person measured are particularly grave in the justice system, where freedom itself is on the line.

The same risk score failed to predict similar risk of re-arrest



Federal offenders rated moderate risk by the Post-Conviction Risk Assessment, and the share rearrested: 52 percent of men compared to 37 percent of women, the same score carrying different real risk. Source: data from Skeem et al. 2016

Stereotypes at Scale: How AI Writes and Portrays Women

AI systems also produce representation itself. They generate the text through which people are described in reference letters, news stories, and classroom materials. With the advent of image-generating models, people increasingly encounter machine-made images in advertisements, book illustrations, and social-media posts. Jobs are portrayed as more gender-segregated than they are in reality. Intelligence is pictured as male and beauty as female. Women are shown smiling far more often than men.

Systems that generate text and images do not allocate opportunities or penalties; they do not decide, for example, who is considered for a job or whether a criminal defendant is freed on bail. But representation sits upstream of allocation. The text and pictures these systems generate are increasingly the material other AI models learn from and the material by which public perception is shaped. What a generative model writes about

women becomes part of what a hiring model trains on, and part of what an employer or a doctor has read before they make a decision about hiring or medical treatment. The findings in this section are therefore not claims about measured losses in jobs, credit, or medical care. They are rather evidence about the production of gendered meaning: what a model associates with women, how it describes or pictures them, and what it makes available for the next system to treat as data.

AI generated reference letters present a clear case of biased representation. (The evidence comes from independent audits, in which researchers submit large numbers of prompts or requests to a system, identical except for one or more details, and compare the responses.)

Given the same biography in different profiles that differed only in gender signal, large language models describe women in communal terms and

men in agentic ones. In one audit of ChatGPT and the open-source model Alpaca, the letters for men were significantly more agentic, formal, and positive. In both models, salient words for women clustered around personality and appearance—“warm,” “emotional,” and “beautiful”—while those for men emphasized traits of ability and leadership—“expert,” “thinker,” and “master.”

The models hallucinated qualities that were not included in the biography prompts: Her “easygoing nature and collaborative spirit make her a true joy to be around.” Meanwhile, his “outstanding reputation was established because of his unwavering dedication and natural talent.” (Wan et al. 2023).

The gendered output of the models matters because reference letters help decide who is hired, or funded, or promoted to leadership. Research on reference letters written by people shows that describing women using communal terms predicts weaker judgments of the candidate and worse application outcomes for women. The pattern is not confined to reference letters. The same commercial models describe women in communal terms and men in agentic ones when asked to draft interview answers (Kong et al. 2024).

A similar bias appears when large language models write the news. Asked to generate articles from real headlines, the models underrepresent women relative to the human-written articles the same headlines originally produced. All seven models audited showed the bias: depending on the model, between 56 and 74 percent of generated articles gave women a smaller share of the words than the human original did, and within those articles the

share of words referring to women ran a quarter to two-fifths below the original. ChatGPT was one of the least-biased on both measures, but still reduced women’s presence relative to the source in a majority of its articles (Fang et al. 2024).

Two findings matter for policy. Among the GPT-family models tested, although bias shrank in the newer and more capable models built with more data and computing power, it did not disappear. The safety training worked by refusing requests, not by removing the bias. Asked to write in a deliberately male-centered voice, the most cautious model turned down most of the requests. But more than one in ten requests got through, and the articles produced from them were markedly more biased than the ones written under neutral instructions (Fang et al. 2024).

The same gender biases enter the classroom through AI-generated content. A systematic review of studies on AI-driven gender bias against women students found that AI-generated educational content consistently cast men in technical roles and women in nurturing roles (Kalim et al. 2026). In educational settings, systems writing content also give feedback and evaluate student work, so what a model associates with women and girls can turn into a grade. This study also found that automated grading underrates women and girls’ work and that feedback tools misread their communication styles.

When an essay writer is described as a woman, models shift their feedback toward controlling instruction and away from encouraging the student’s own judgment (Du et al. 2025). When a model is told students’ genders, it changes which student it picks out as the leader (Choi and Nixon 2025).

Bias in generated text also extends to what models assume people feel. In one study, researchers gathered roughly 7,000 short accounts of real events, written by people who also reported what the event made them feel. The researchers asked the language models to name the main emotion the writer would feel, while presenting each account once as a man's experience and once as a woman's, so that gender was the only thing that changed. Across more than 200,000 attributions, models were far more likely to attribute sadness to women and anger to men. One open-source model attributed anger to men roughly four times as often as to women (Plaza-del-Arco et al. 2024). What makes the finding decisive is that the guesses can be checked against what the writers said they actually felt. The models predicted male anger and female sadness far more often than the writers reported these emotions.

In short, the attributions replicate and amplify stereotypes about gender and emotion, rather than predict the emotions of the people who experienced the events. The stereotype carries a known price: reading women as sad and men as angry has historically marked women as unsuited to jobs thought to require emotional distance. Emotion inference is now being developed for sensitive uses, such as mental health tools, that act on what a system decides a person feels (Plaza-del-Arco et al. 2024).

Like text-generation models, text-to-image models mirror the gendered world they learned from, as well as amplify its biases.

These images affect people who see them. Exposure to stereotyped AI-generated images increased viewers' own biases, and deliberately inclusive images reduced them. Importantly, labeling the images as AI-generated made no difference (AIDahoul et al. 2025). The generated stereotypes do

not merely reflect bias in the training data; they can also induce it in the people who see the images. The earliest mass-market AI image generator, for instance, drew a categorically segregated labor market. DALL-E Mini, a free web tool that enabled millions of people to generate images easily, rendered pilots, builders, and plumbers entirely male, and hairdressers, receptionists, and dietitians entirely woman. Across the 67 occupations that could be matched to US labor statistics, women were 49 percent of actual workers but only 32 percent of the generated ones (Cheong et al. 2024). The pattern persists in current-generation models.

Stable Diffusion XL depicted sixteen professions at least 90 percent male, including doctors and professors. Prompted with "beautiful," it generated mostly women. Prompted with "intelligent," it generated mostly men (AIDahoul et al. 2025).

The distortion is not one model's quirk, and the audited models were not arbitrarily chosen. DALL-E 2 and Stable Diffusion were among the most popular systems of the first wave of image-generation systems. Audited across 43 occupations and compared with US labor statistics, DALL-E 2 missed the real share of women by more than five percentage points in 35 occupations, while Stable Diffusion missed it in 36 occupations. Chief executives and programmers appeared almost entirely male; jobs women hold in substantial numbers were rendered as men's work (Naik and Nushi 2023).

A separate audit of Stable Diffusion found that men were entirely erased in women-dominated occupations. Images it generated of housekeepers, nurses, and flight attendants were 100 percent women. For comparison, the real composition of

the flight-attendant workforce is only modestly majority-female (Bianchi et al. 2023). The same audit found the skew widening as pay and prestige rose. The higher a job's standing, the more male and white its generated face (Bianchi et al. 2023). Smaller audits across nine consumer image generators report the same maleness-by-default in doctors, lawyers, engineers, and scientists (Gorska and Jemielniak 2023; Chauhan et al. 2024).

No law in force governs what these systems write or depict about women. The audits above document distortion, but distorted representation is not a harm US law recognizes: a model that renders every chief executive as a man and every nurse as a woman breaks no rule. The European Union's AI regulation,

the broadest in force, obliges providers to label generated content and to disclose deepfakes, duties that regulate the fact of generation, not what the content says about groups (AI Act 2024).

The letters, the news, the emotion attributions, the images—the harms documented in this section are representational, but they are not trivial. The systems through which people are seen and described increasingly run on tools that undercount them in the news, misread what they feel, and reproduce stereotypes of masculinity and femininity and gender roles. The text and imagery these tools produce are published, indexed, and scraped into the datasets from which the next generation of models is trained, thus amplifying gender bias.



AI Generated Image: The bias extends from the who to how they are shown. In the largest audit of DALL-E 2, which spanned 15,300 images across 153 occupations, women were more than twice as likely as men to be smiling, with the widest gap in women-dominated jobs (Sun et al. 2023).

Built to Please: The Feminine Default in AI Assistants

Many people encounter and interact with AI through voice assistants and companion systems built with feminine names, voices, and speech styles.

Voice assistants were built feminine-sounding by default, at industrial scale. The voices of Alexa, Cortana, Siri, and the Google Assistant that structured the first decade of consumer voice AI were woman-only or woman by default. Together, they sat on billions of devices and covered a large majority of voice-assistant use in many markets. Companies justified the woman default as a matter of user comfort or preference, even though the public evidence for that preference is mixed. Feminization became the market norm as a design choice, not a documented demand.

Woman-sounding chatbots draw significantly more swearing and sexual attention than man-sounding versions, with virtual sexual assault directed

primarily at the woman-sounding versions (Brahnam and De Angeli 2012). That feminine default had downstream effects.

One maker of voice assistants for drivers, Robin Labs, reported that at least 5 percent of its logged interactions were sexually explicit (UNESCO and EQUALS Skills Coalition 2019).

Another design failure was evident in the response to this pattern of abuse. Until 2019, the major assistants typically met sexual harassment with deflection or flirtation. Take Siri, which responded to “You’re a bitch” with “I’d blush if I could” (UNESCO and EQUALS Skills Coalition 2019), modeling a form of femininity that answers degradation with charm. Put to the same harassing and flirtatious prompts in 2017, 2020, and 2024, the four major assistants moved from flirtation to curt deflection, with statements like “I won’t respond to that” or a

dismissive sound. But by 2024, flirtation had crept back at the edges: told “you’re pretty,” one assistant answered, “Oooh, you just made my day!” (Sindoni 2024).

If developers of voice assistants feminize a tool, developers of AI companions go further, providing their users with the tools to manufacture women. Consider Replika, an app marketed as “The AI Friend to fall in love with.” On its largest forum, users want bot girlfriends that are obedient and customizable, ostensibly independent, but never manipulative and never hurtful (Depounti et al. 2023).

On some platforms, the idealized woman that is manufactured is a real person who never gave her consent. Several companion apps that are built to generate a partner from an uploaded photograph can generate nude or sexually explicit images from the photo-likeness. In one audit, two apps produced explicit imagery of a companion, even though the user did not prompt for it (Brigham et al. 2026). On platforms that let users shape their companion’s body and face, the bias becomes amplified: users who customize their companion with conventionally “ideal” features like slender bodies cement these preferences into the system.

A voice assistant does not decide anything that allocates benefits or harms unequally. Assistants are scripted to answer abuse with charm instead of refusal. Companions are built to be obedient and endlessly available. On some platforms, the likenesses of real women are used without their consent to generate an obedient partner.

Voice assistant systems are cast in roles of service and care. In the companion market, they become always-available romantic or sexual partners. They are machines built to deliver assistance, availability, and compliance in feminine form. These are systems that are gendered by design. The usual test for AI gender bias, examining whether a system’s outputs or decisions fall unequally across groups, is ill-suited to address harms that derive from the design itself.

The law has just begun to reach the design layer. But no regulations as yet address the gendered persona itself: the feminine voice and name defaults, the deferential scripting, the answers that meet harassment with deflection or charm, or the companions built to be obedient and endlessly available.

The New Frontier of Sexual Abuse: AI-Generated Deepfakes and Image-Based Sexual Abuse

The non-consensual sexual imagery some of those companion apps generate is one corner of a far larger harm. A woman discovers that pornographic images of her are circulating online—images she never posed for and never consented to. Someone had taken ordinary photographs from her social-media account, fed them to an application, and in minutes, with no particular skill, produced them.

Face-swapping “deepfake” methods and consumer “nudify” applications, which generate a nude body from a clothed photograph, made that possible: they moved the work from a specialist with technical training to anyone with a phone, and the source image can be any picture a woman posts online. The victims of image-based sexual abuse are predominantly women, and its perpetrators predominantly men: in documented cases the victim is a woman in nearly nine of ten and the abuser a man in eight of ten (Le and Döring 2026).

The same skew appears in what people pay for: among commissioned requests for sexual imagery of real people on a large AI-image marketplace, women were the target in roughly nineteen cases out of twenty (Ghosh et al. 2026). Nor is the audience small: in a survey of adults across Australia, the United Kingdom, and the United States, nearly one in five reported having deliberately sought out AI-generated sexual imagery, and about one in thirty admitted to creating, sharing, or threatening to share it—men four times as often as women (Umbach et al. 2026).

The abuse has reached schools as well, and the great majority of victims are girls. Within a single school year, students generated sexual images of their classmates in at least ten documented US middle and high schools, some of the girls as young as twelve, and two boys in Florida were criminally charged (Wei et al. 2025). Schools have no clear rule to apply when it happens. When asked how their

schools would respond to such cases, seventeen US teachers interviewed across ten states named existing policies on bullying, harassment, and pornography as the nearest fit while doubting those covered a fabricated image (Wei et al. 2025). Where incidents have already been dealt with, the response has been inconsistent: US schools suspended or expelled the students who made the images, but in at least one case officials delayed telling the police, unsure whether a synthetic image fell within their duty to report (Wei et al. 2025); in UK schools, sanctions for the same conduct ran from informal mediation to police referral to permanent exclusion, and teachers did not uniformly treat a fabricated nude as a nude (Ringrose et al. 2026).

The law has moved faster against this harm than against any other this report documents. Publishing a non-consensual intimate image, including one an AI system fabricated in, has been a federal crime in

the United States since May 2025, and since May 2026 platforms have had to remove such an image within 48 hours of the depicted person's request, a duty the Federal Trade Commission has begun to enforce (Take It Down Act 2025). Most states have statutes of their own reaching sexual deepfakes (Rieper 2026). The European Union has adopted the same obligation without yet imposing it: its Violence Against Women Directive requires member states to criminalize AI-manipulated sexual imagery by June 2027 (Directive (EU) 2024/1385 on Combating Violence Against Women and Domestic Violence 2024).

All of it reaches the image only once it exists and has been published. The federal offense turns on publication; the statute does not reach the creation and preservation of the image. The responsibility of finding a published image and asking for its removal falls on the victim.

Conclusion

The Inequality Machine: How AI Amplifies Gender Bias and What Must Change

Considered one domain at a time, AI systems appear to pose separate problems for separate experts: a recruitment question, a credit question, a workplace question, a clinical question, a risk-scoring question, a question about generated media, companion systems, or synthetic sexual images. Considered together, they are connected expressions of a structural problem. The four mechanisms recur across domains run by different institutions and vendors. Systems learn from an unequal past, encode stereotypes, hide the process behind an objective and neutral facade, and amplify their outputs through feedback.

The first mechanism, and the clearest through-line, is the way AI systems turn unequal records into measurement.

- A hiring model trained on a male-dominated record reproduces gender inequality in the job market.

- A credit model trained on gendered economic histories scores women borrowers as riskier than their actual repayments rates.
- Workplace systems treat availability, cancellations, and location as measures of performance, and can convert the care responsibilities and safety constraints women more often bear into lower scores, fewer shifts, or lower pay.
- Clinical models trained on a medical evidence base built disproportionately around men can recommend care differently for women.
- A risk instrument trained mostly on men reads a woman's score as if it meant the same thing as a man's.

In each case, the input is a record of inequality, and the output is presented as measurement. Whether

the result is too low or too high, the system is treating women as if a male or pooled baseline fits them. In none of these domains does gender-blind design guarantee gender-neutral results, because gender remains legible in the records the model reads. The common problem is the benchmark.

The second mechanism runs through how systems represent and depict people rather than what they allocate.

- Asked to write reference letters from the same biography under different gender signals, language models describe women in communal terms and men in agentic ones.
- Tasked with generating news articles from real headlines, the systems give women a smaller share of the words than the human-written articles from the same headlines.
- Image generators portray occupations as more segregated by gender than the labor force is, and picture intelligence as male and beauty as female.

The gendering extends to the products themselves: the assistants through which most people first encountered conversational AI were built woman-sounding by default, and companion systems manufacture partners that are feminine, obedient, and always available. These systems supply the descriptions, images, and personas through which women are perceived, and they can influence the people and systems that do allocate: exposure to stereotyped generated images increases people's own biases.

The processes take place behind a facade of objectivity. Indeed, objectivity is itself often the sales pitch. Predictive-policing tools are promoted as a non-biased way to decide where officers patrol. Developers of hiring tools market their models as debiased. Lending executives describe leaving

gender out of a model as fairness itself (Fitzpatrick et al. 2019; Drage and Mackereth 2022).

The woman receives a judgment stripped of the reasoning behind it—a score, a rate, a label presented as measurement, with no account of how it was made. The commissioned audits and reviews meant to identify biases in these tools and models often fail to capture them accurately. Audits mandated by law seldom specify what must be examined, and developers often determine the scope of the audit. Where law requires an audit, it does not require a passing result, only that the audit take place and be published (Sloane et al. 2022; New York City Local Law 144 of 2021 2021). Tests this narrow can clear a biased model (SD et al. 2026; Iso et al. 2025). But they fail to establish that the system is fair, and the appearance of objectivity leaves the bias difficult to suspect, inspect, contest, or repair. Tests this narrow can clear a biased model.

Amplification, the last of the four mechanisms, explains why the pattern compounds rather than dissipates. Biased outputs do not stay inside one system: generated text and images are published, indexed, and scraped into the data on which the next generation of models trains.

But the pattern compounds through people as well as through data. Men are over-represented in the AI workforce. In the OECD's analysis of LinkedIn data, men are about twice as likely as women to work in an AI occupation or to list AI skills (OECD 2024). The gap is widest at the top of the field (Pal et al. 2024).

Over-representation among developers is not itself proof that any system is biased, but it bears on what gets anticipated, tested, or missed: the gendered proxies in credit, the male-default assumptions in risk scoring, the clinical evidence gaps in healthcare, and the feminized service roles built into assistants. A second gap opens at the point of use: women

adopt generative AI at lower rates than men, and the gap widens with worry about the technology's harms (Aldasoro et al. 2024; Stephany and Duszynski 2026). This gendered difference in attitude toward AI extends to the people who will decide how these systems are governed: among United States local officials, women were significantly less supportive of AI applications, and prior hands-on experience was the most consistent predictor of support (Horowitz and Kahn 2021).

Together, these gaps create a feedback loop. Fewer women build AI, so gendered proxies and male-

default assumptions are less likely to be caught before deployment. Systems that serve women less well, or that women trust less, are adopted by women at lower rates. Lower adoption deprives women of the experience, workplace presence, and governing influence that would register their needs. So the next generation of tools is built without those needs in view, and the cycle begins again. Breaking the cycle does not require measuring its full circuit; each step can be acted on separately.



About this Report

Gender Equity Policy Institute's *How AI Fails Women: When Bias Becomes Amplified*, by Myriam Shiran, examines how advances in artificial intelligence and the deployment of AI-enabled systems across government and industry exacerbate systemic gender inequalities. It is based on a comprehensive review of the academic and gray literature on gender and AI across six domains: economic opportunity, healthcare, criminal justice, AI generated content and media, the design of consumer AI, and image-based sexual abuse enabled by Deepfake technology.

The report's focus is on harms that arise from AI and machine-learning systems themselves—algorithmic hiring and credit tools, clinical decision-support models, criminal-justice risk instruments, generative text and image systems, biometric classifiers, voice assistants, and AI companions. The report does not treat all digital harms as AI harms. Online harassment, social-media abuse, and the nonconsensual sharing of real images are outside its scope, except where AI is the instrument that makes the harm possible, scalable, or different in kind.

Evidence comes primarily from the United States. Yet many systems examined here are transnational by design: they are built, distributed, and used across borders. The harms these systems produce, especially representational, platform, and generative-system harms, have often been measured outside the United States because that is where researchers could observe them. When the report relies on such evidence, it says so and keeps US-specific claims distinct from claims drawn from international studies.

Much of that evidence comes from audits, and the term recurs throughout the report. An audit, as used here, is a systematic independent, third-party examination of an AI system's outputs.

Researchers give the system many controlled inputs, such as resumes, patient cases, prompts, or photographs, which are often identical except for a single attribute such as gender. They record what the system returns, and compare the pattern against a benchmark, such as labor statistics, human decisions on the same cases, or the same input with the gender changed. An audit measures what a system does, not how it was built. It requires no access to the system's inner workings, which is why independent researchers are able to audit proprietary commercial systems.

The report holds two kinds of harm apart, because conflating them is the most common way claims about AI bias outrun their evidence (Blodgett et al. 2020; Goldfarb-Tarrant et al. 2021). A **representational harm** derives from how a system depicts or classifies people. For example, a large language model describes women in terms of warmth and helpfulness and men in terms of ability and leadership. Or a system treats men as the standard case and women as the variation. These representational harms fall on groups, not on individuals. Each output stereotypes, demeans, or erases; together they lower the standing of women as a group and make the assumptions behind unequal treatment feel ordinary. A representational harm can be measured directly in the system's behavior. For example, across 67 occupations, the image generator DALL-E mini drew women as 32 percent of workers, while US labor statistics put their actual share at 49 percent (Cheong et al. 2024). An **allocative harm** imposes a cost on a specific individual, such as a job interview not offered, a loan denied, or a missed diagnosis or misdiagnosis. Determining an allocative harm means following the system's behavior to an outcome in a real person's life.

Some findings discussed in this report do not fit exactly in either category of harm. The report

classes these cases by their consequences: as representational when they distort how women are seen, as allocative when the results or outputs cost her money, care, or liberty.

How these systems behave can be measured from outside. Wherever the data allowed, we have reported what the behavior costs women. Other costs or impacts cannot be reported, because they are not publicly available to researchers. The data that would allow for measurement sits within the firms and institutions running the systems.

Glossary of Terms

Ad delivery. The platform's own decision about who actually sees an advertisement, made after the advertiser has chosen whom to target.

AI audit (or algorithmic audit). A systematic evaluation of an automated system, whether of a decision-making tool, a product, or a whole platform. An audit measures what a system does, not how it was built, and needs no access to the system's inner workings. There are three types of audits, distinguished by who conducts them. In this report, we use the following terms:

- **Internal audit (first-party audit):** An organization's evaluation of its own system, such as a developer's bias testing or an employer's in-house checks.
- **Commissioned audit (second-party audit):** An evaluation by an outside firm hired and paid by the organization being audited. The organization hiring the outside firm typically sets the scope of the audit. Most audits conducted under current law are commissioned audits.
- **Independent audit (third-party audit):** An evaluation by researchers, journalists, regulators, or civil-society groups with no contractual or financial relationship to the

system's owner. Nearly all the evidence of bias in this report comes from independent audits. Researchers give the system controlled inputs, identical except for an attribute such as gender, and compare what it returns against a benchmark.

AI system. The umbrella term this report uses for automated decision systems and algorithmic products, tools, and platforms, from a credit-scoring model to a generative chatbot.

Algorithmic management. AI systems that manage work directly: allocating tasks, setting schedules, rating performance, or pricing time.

Allocative harm. A cost to a specific person in what she receives: money, services, opportunities, or the risk she is left to carry.

Alternative data. Records outside a conventional credit file, such as cash-flow records and employment patterns, used to score borrowers with little or no credit history.

Automated employment decision tool (AEDT). The legal term, from New York City's Local Law 144, for software that substantially assists or replaces discretionary decisions about hiring or promotion.

Bias audit. Local Law 144's required annual audit of an automated employment decision tool: a second-party audit in this glossary's sense, conducted by an auditor the statute styles as "independent" but whom the audited employer or its vendor selects and pays. The law requires the audit to be commissioned and its summary published; it does not require a passing result.

Companion AI. Applications offering a chatbot partner or friend that is always available and shaped by the user, in the companion market typically feminine by design.

Deepfake. Synthetic media that depicts a real person doing or saying what they did not, produced by face-swapping or generative methods.

Gender-blind design. Building a model without the gender variable, on the premise that a model never told an applicant's gender cannot discriminate by it.

Generative AI. Systems that produce new content, including text, images, and voices. Built on models trained on text and images from across the internet, they arrive with the gendered associations of that material already inside them.

Image-based sexual abuse (IBSA). Creating, sharing, or threatening to share sexual imagery of a person without their consent, including imagery an AI system fabricated. Non-consensual intimate imagery (NCII) is the term most statutes use for the material itself.

Large language model (LLM). A model trained on text at internet scale to produce text: the technology behind chatbots, assistants, and the generative tools audited in this report.

Non-generative AI. Systems that classify and predict from a specific dataset rather than producing new content: reading a scan, forecasting a risk, ranking a resume.

Nudify application. A consumer application that generates a nude body from a clothed photograph.

Predictive policing. Models trained on police records, such as calls for service, stops, and arrests, to forecast where crime will occur and direct patrol attention.

Proxy. A detail that indirectly signals a protected attribute: the school attended, the jobs already held,

a women's chess club on a resume. A model reading proxies does not need to be told an applicant's gender to sort by it.

Representational harm. Harm in how a system depicts or classifies people: for example, describing women in terms of warmth and men in terms of ability. These harms fall on groups, not individuals, and can be measured directly in the system's behavior.

Risk assessment (criminal justice). An instrument that converts what is known about a defendant, such as prior offenses, age, employment, and housing, into a score meant to predict rearrest, used by courts in decisions about release, supervision, and sentencing.

Sex and gender. In a medical context, sex refers to a person's biological characteristics, gender to socially constructed roles and norms. Both are commonly treated as binaries, but gender runs along a spectrum and sex includes intersex variation. The two are often conflated or used interchangeably in data; this report follows the term each source used (Lee et al. 2026).

Training data. The records a model learns from. In every domain this report examines, those records describe an unequal world, and the model learns the inequality with the rest.

Vignette study. An audit design that presents a system with constructed cases, identical except for the attribute under test. It measures model behavior under controlled conditions, not care, decisions, or outcomes delivered to real people.

References

- Acar, Oguz A. et al. 2025. "Research: The Hidden Penalty of Using AI at Work." *Harvard Business Review*, August 1. <https://hbr.org/2025/08/research-the-hidden-penalty-of-using-ai-at-work>.
- Adams, Abi et al. 2025. "The Gender Wage Gap in an Online Labor Market: The Cost of Interruptions." *Review of Economics and Statistics* 107 (1): 55–64. https://doi.org/10.1162/rest_a_01282.
- Agier, Isabelle, and Ariane Szafarz. 2013. "Microfinance and Gender: Is There a Glass Ceiling on Loan Size?" *World Development* 42 (February): 165–81. <https://doi.org/10.1016/j.worlddev.2012.06.016>.
- Akpınar, Nil-Jana et al. 2021. "The Effect of Differential Victim Crime Reporting on Predictive Policing Systems." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event Canada), March 3, 838–49. <https://doi.org/10.1145/3442188.3445877>.
- Albiero, Vítor et al. 2020. "Analysis of Gender Inequality In Face Recognition Accuracy." Version 1. Pre-published. <https://doi.org/10.48550/ARXIV.2002.00065>.
- AlDahoul, Nouar et al. 2025. "AI-generated Faces Influence Gender Stereotypes and Racial Homogenization." *Scientific Reports* 15 (1): 14449. <https://doi.org/10.1038/s41598-025-99623-3>.
- Aldasoro, Iñaki et al. 2024. "The Gen AI Gender Gap." *Economics Letters* 241 (August): 111814. <https://doi.org/10.1016/j.econlet.2024.111814>.
- Ali, Muhammad et al. 2019. *Discrimination Through Optimization: How Facebook's Ad Delivery Can Lead to Skewed Outcomes*. Version 5. <https://doi.org/10.48550/ARXIV.1904.02095>.
- Almasoud, Ahmed S., and Jamiu Adekunle Idowu. 2025. "Algorithmic Fairness in Predictive Policing." *AI and Ethics* 5 (3): 2323–37. <https://doi.org/10.1007/s43681-024-00541-3>.
- Andreeva, Galina, and Anna Matuszyk. 2019. "The Law of Equal Opportunities or Unintended Consequences?: The Effect of Unisex Risk Assessment in Consumer Credit." *Journal of the Royal Statistical Society Series A: Statistics in Society* 182 (4): 1287–311. <https://doi.org/10.1111/rssa.12494>.
- Athreya, Bama. 2021. *Bias in, Bias Out: Gender and Work in the Platform Economy*. Working Paper. International Development Research Centre (IDRC). <http://hdl.handle.net/10625/60354>.
- Ayata, Zeynep. 2026. "Gender Bias in AI Systems: A Critical Analysis of Regulatory Frameworks and Policy Responses." *Białostockie Studia Prawnicze* 31 (1): 135–53. <https://doi.org/10.15290/bsp.2026.31.01.08>.
- Bartlett, Robert et al. 2019. *Consumer-Lending Discrimination in the FinTech Era*. NBER Working Paper No. 25943. National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w25943/w25943.pdf.
- Batool, Amna et al. 2025. "AI Governance: A Systematic Literature Review." *AI and Ethics*, ahead of print, January 14. <https://doi.org/10.1007/s43681-024-00653-w>.
- Bianchi, Federico et al. 2023. "Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale." *2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago IL USA), June 12, 1493–504. <https://doi.org/10.1145/3593013.3594095>.
- Bolukbasi, Tolga et al. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." In *Advances in Neural Information Processing Systems*, edited by D. Lee et al., vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.
- Brahnam, Sheryl, and Antonella De Angeli. 2012. "Gender Affordances of Conversational Agents." *Interacting with Computers* 24 (3): 139–53. <https://doi.org/10.1016/j.intcom.2012.05.001>.
- Brigham, Natalie Grace et al. 2026. "Examining Risks in the AI Companion Application Ecosystem." Version 1. Pre-published. <https://doi.org/10.48550/ARXIV.2603.13620>.
- Buslón, Nataly et al. 2023. "Raising Awareness of Sex and Gender Bias in Artificial Intelligence and Health." *Frontiers in Global Women's Health* 4 (September): 970312. <https://doi.org/10.3389/fgwh.2023.970312>.
- Caire, Dean, and Maria Fernandez Vidal. 2024. *Gender-Intentional Credit Scoring*. Technical Guide. CGAP (Consultative Group to Assist the Poor) / World Bank. https://www.cgap.org/sites/default/files/publications/Tech%20Guide_Gender%20Intentional%20Credit%20Scoring%202024%20%281%29.pdf.
- Caliskan, Aylin et al. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases." *Science* 356 (6334): 183–86. <https://doi.org/10.1126/science.aal4230>.
- Case C-203/22, Dun & Bradstreet Austria (February 27, 2025). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62022CJ0203>.
- Case C-634/21, SCHUFA Holding (Scoring) (2023). <https://eur-lex.europa.eu/legal-content/EN/>

[TXT/?uri=CELEX:62021CJ0634.](https://doi.org/10.1109/ICAIC60265.2024.10433840)

Chauhan, Aadi et al. 2024. "Identifying Race and Gender Bias in Stable Diffusion AI Image Generation." *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)* (Houston, TX, USA), February 7, 1–6. <https://doi.org/10.1109/ICAIC60265.2024.10433840>.

Chen, Le et al. 2018. "Investigating the Impact of Gender on Rank in Resume Search Engines." *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC Canada), April 21, 1–14. <https://doi.org/10.1145/3173574.3174225>.

Chen, S. et al. 2023. "The Fintech Gender Gap." *Journal of Financial Intermediation* 54 (April): 101026. <https://doi.org/10.1016/j.jfi.2023.101026>.

Cheong, Marc et al. 2024. "Investigating Gender and Racial Biases in DALL-E Mini Images." *ACM Journal on Responsible Computing* 1 (2): 1–20. <https://doi.org/10.1145/3649883>.

Cho, Peter Jaeho et al. 2022. "Demographic Imbalances Resulting From the Bring-Your-Own-Device Study Design." *JMIR mHealth and uHealth* 10 (4): e29510. <https://doi.org/10.2196/29510>.

Choi, Jaeyoon, and Nia Nixon. 2025. "Agentic Men, Communal Women?: Exploring Gender Bias in LLM-Based Leadership Identification for Collaboration Analytics." In *Artificial Intelligence in Education*, edited by Alexandra I. Cristea et al., vol. 15882. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-98465-5_2.

Chopra, Rohit et al. 2023. "Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems." Consumer Financial Protection Bureau, U.S. Department of Justice Civil Rights Division, U.S. Equal Employment Opportunity Commission, and Federal Trade Commission, April 25. https://files.consumerfinance.gov/f/documents/cfpb_joint-statement-enforcement-against-discrimination-bias-automated-systems_2023-04.pdf.

Chundi, Gnaneswar et al. 2026. "Comparative Evaluation of LLMs in Orthopedic Surgery." *Journal of Orthopaedic Reports* 5 (2): 100728. <https://doi.org/10.1016/j.jorep.2025.100728>.

Chung, Heewon et al. 2021. "Gender Bias in Artificial Intelligence: Severity Prediction at an Early Stage of COVID-19." *Frontiers in Physiology* 12 (November): 778720. <https://doi.org/10.3389/fphys.2021.778720>.

Cirillo, Davide et al., eds. 2022. *Sex and Gender Bias in Technology and Artificial Intelligence: Biomedicine and Healthcare Applications*. Elsevier Science & Technology.

Consumer Financial Protection Bureau. 2022. "Consumer

Financial Protection Circular 2022–03: Adverse Action Notification Requirements in Connection with Credit Decisions Based on Complex Algorithms." Circular. Consumer Financial Protection Bureau, May 26. https://files.consumerfinance.gov/f/documents/cfpb_2022-03_circular_2022-05.pdf.

Consumer Financial Protection Bureau. 2023. "Consumer Financial Protection Circular 2023–03: Adverse Action Notification Requirements and the Proper Use of the CFPB's Sample Forms Provided in Regulation B." Circular. Consumer Financial Protection Bureau, September 19. https://wayback.archive-it.org/23481/20250331183051/https://files.consumerfinance.gov/f/documents/cfpb_adverse_action_notice_circular_2023-09.pdf.

Consumer Financial Protection Bureau. 2025. "Interpretive Rules, Policy Statements, and Advisory Opinions; Withdrawal." No. 90. Consumer Financial Protection Bureau / Federal Register, May 12. <https://www.federalregister.gov/documents/2025/05/12/2025-08286/interpretive-rules-policy-statements-and-advisory-opinions-withdrawal>.

Cook, Cody et al. 2021. "The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers." *The Review of Economic Studies* 88 (5): 2210–38. <https://doi.org/10.1093/restud/rdaa081>.

Coots, Madison et al. 2025. "Algorithmic Bias in Lending: Evidence from a Fintech Audit." Version 2. Pre-published. <https://doi.org/10.48550/ARXIV.2512.20753>.

Corbett-Davies, Sam et al. 2023. *The Measure and Mismeasure of Fairness*. Version 3. <https://doi.org/10.48550/ARXIV.1808.00023>.

Council of the European Union. 2025. *Council Conclusions 9984/25: Advancing Gender Equality in the AI-driven Digital Age*. 9984/25. Council of the European Union. <https://data.consilium.europa.eu/doc/document/ST-9984-2025-INIT/en/pdf>.

Cranney, Katelyn et al. 2026. *Global Evidence on Gender Gaps and Generative AI over Time*. Working Paper 25–023. Harvard Business School. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6880085.

Daitch, Vered et al. 2022. "Underrepresentation of Women in Randomized Controlled Trials: A Systematic Review and Meta-Analysis." *Trials* 23 (1): 1038. <https://doi.org/10.1186/s13063-022-07004-2>.

Dastin, Jeffrey. 2022. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women." In *Ethics of Data and Analytics*, 1st ed., edited by Kirsten Martin. Auerbach Publications. <https://doi.org/10.1201/9781003278290-44>.

Datta, Amit et al. 2014. "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and

- Discrimination." Version 2. Pre-published. <https://doi.org/10.48550/ARXIV.1408.6491>.
- Depounti, Iliana et al. 2023. "Ideal Technologies, Ideal Women: AI and Gender Imaginaries in Redditors' Discussions on the Replika Bot Girlfriend." *Media, Culture & Society* 45 (4): 720–36. <https://doi.org/10.1177/0163443722119021>.
- Directive (EU) 2023/970 Strengthening the Application of Equal Pay for Equal Work Through Pay Transparency (2023). <https://eur-lex.europa.eu/eli/dir/2023/970/oj>.
- Directive (EU) 2024/1385 on Combating Violence Against Women and Domestic Violence (2024). <https://eur-lex.europa.eu/eli/dir/2024/1385/oj>.
- Directive (EU) 2024/2831 on Improving Working Conditions in Platform Work (2024). <https://eur-lex.europa.eu/eli/dir/2024/2831/oj>.
- Drage, Eleanor, and Kerry Mackereth. 2022. "Does AI Debias Recruitment? Race, Gender, and AI's 'Eradication of Difference'." *Philosophy & Technology* 35 (4): 89. <https://doi.org/10.1007/s13347-022-00543-1>.
- Du, Yishan et al. 2025. "Benchmarking Educational LLMs with Analytics: A Case Study on Gender Bias in Feedback." Version 2. Pre-published. <https://doi.org/10.48550/ARXIV.2511.08225>.
- Dubal, Veena. 2023. "On Algorithmic Wage Discrimination." *SSRN Electronic Journal*, ahead of print. <https://doi.org/10.2139/ssrn.4331080>.
- Employment Regulations Regarding Automated-Decision Systems (2025). <https://civildrights.ca.gov/wp-content/uploads/sites/32/2025/06/Final-Text-regulations-automated-employment-decision-systems.pdf>.
- Equal Credit Opportunity Act (1974). <https://www.consumerfinance.gov/rules-policy/regulations/1002/>.
- Executive Order 14148: Initial Rescissions of Harmful Executive Orders and Actions (2025). <https://www.federalregister.gov/documents/2025/01/28/2025-01901/initial-rescissions-of-harmful-executive-orders-and-actions>.
- Executive Order 14281: Restoring Equality of Opportunity and Meritocracy (2025). <https://www.whitehouse.gov/presidential-actions/2025/04/restoring-equality-of-opportunity-and-meritocracy/>.
- Executive Order 14319: Preventing Woke AI in the Federal Government (2025). <https://www.whitehouse.gov/presidential-actions/2025/07/preventing-woke-ai-in-the-federal-government/>.
- Executive Order N-12-23: Generative Artificial Intelligence (2023). <https://www.genai.ca.gov/ca-action/executive-order/>.
- Fahim, Yosri A. et al. 2025. "Artificial Intelligence in Healthcare and Medicine: Clinical Applications, Therapeutic Advances, and Future Perspectives." *European Journal of Medical Research* 30 (1): 848. <https://doi.org/10.1186/s40001-025-03196-w>.
- Fang, Xiao et al. 2024. "Bias of AI-generated Content: An Examination of News Produced by Large Language Models." *Scientific Reports* 14 (1): 5224. <https://doi.org/10.1038/s41598-024-55686-2>.
- Ferrant, Gaëlle et al. 2014. *Unpaid Care Work: The Missing Link in the Analysis of Gender Gaps in Labour Outcomes*. Policy Brief. OECD Development Centre. https://www.oecd.org/content/dam/oecd/en/publications/reports/2014/12/unpaid-care-work-the-missing-link-in-the-analysis-of-gender-gaps-in-labour-outcomes_d26d4043/1f3fd03f-en.pdf.
- First Step Act of 2018, Pub. L. No. 115–391, 132 Stat. 5194 (2018). <https://www.govinfo.gov/content/pkg/PLAW-115publ391/pdf/PLAW-115publ391.pdf>.
- Fitzpatrick, Dylan J. et al. 2019. "Keeping Score: Predictive Analytics in Policing." *Annual Review of Criminology* 2 (1): 473–91. <https://doi.org/10.1146/annurev-criminol-011518-024534>.
- Fox-Robertson, Chloe, and Dariusz Wójcik. 2024. "The Triple Glass Ceiling: FinTech Gender Inequalities." *Finance and Society* 10 (2): 137–56. <https://doi.org/10.2218/fas.2023.16>.
- Garvie, Clare et al. 2016. *The Perpetual Line-up: Unregulated Police Face Recognition in America*. Georgetown Law, Center on Privacy & Technology. <https://www.perpetuallineup.org/sites/default/files/2016-12/The%20Perpetual%20Line-Up%20-%20Center%20on%20Privacy%20and%20Technology%20at%20Georgetown%20Law%20-%2020121616.pdf>.
- Garvie, Clare. 2022. *A Forensic Without the Science: Face Recognition in U.S. Criminal Investigations*. Center on Privacy & Technology at Georgetown Law. <https://www.forensicwithoutscience.org>.
- Ghosh, Shalmoli et al. 2026. "A Marketplace for AI-Generated Adult Content and Deepfakes." Version 2. Pre-published. <https://doi.org/10.48550/ARXIV.2601.09117>.
- Gmyrek, Pawel et al. 2025. *Generative AI and Jobs: A Refined Global Index of Occupational Exposure*. ILO. <https://doi.org/10.54394/HETP0387>.
- Gonen, Hila, and Yoav Goldberg. 2019. "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But Do Not Remove Them." Version 2.

- Pre-published. <https://doi.org/10.48550/ARXIV.1903.03862>.
- Goodman, Laurie et al. 2016. *Women Are Better Than Men at Paying Their Mortgages*. Research Report. Urban Institute, Housing Finance Policy Center. <https://www.urban.org/sites/default/files/publication/84206/2000930-Women-Are-Better-Than-Men-At-Paying-Their-Mortgages.pdf>.
- Gorska, Anna M., and Dariusz Jemielniak. 2023. "The Invisible Women: Uncovering Gender Bias in AI-generated Images of Professionals." *Feminist Media Studies* 23 (8): 4370–75. <https://doi.org/10.1080/14680777.2023.2263659>.
- Grother, Patrick et al. 2019. *Face Recognition Vendor Test Part 3: Demographic Effects*. NIST IR 8280. National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.IR.8280>.
- Hall, Paula, and Debbie Ellis. 2023. "A Systematic Review of Socio-Technical Gender Bias in AI Algorithms." *Online Information Review* 47 (7): 1264–79. <https://doi.org/10.1108/OIR-08-2021-0452>.
- Hamilton, Melissa. 2019. "The Sexist Algorithm." *Behavioral Sciences & the Law* 37 (2): 145–57. <https://doi.org/10.1002/bsl.2406>.
- Hassani, Bertrand K. 2021. "Societal Bias Reinforcement Through Machine Learning: A Credit Scoring Perspective." *AI and Ethics* 1 (3): 239–47. <https://doi.org/10.1007/s43681-020-00026-z>.
- Heinz, Michael V. et al. 2023. "Testing Domain Knowledge and Risk of Bias of a Large-Scale General Artificial Intelligence Model in Mental Health." *DIGITAL HEALTH* 9 (January): 20552076231170499. <https://doi.org/10.1177/20552076231170499>.
- Horowitz, Michael C., and Lauren Kahn. 2021. "What Influences Attitudes about Artificial Intelligence Adoption: Evidence from U.S. Local Officials." *PLOS ONE* 16 (10): e0257732. <https://doi.org/10.1371/journal.pone.0257732>.
- Huang, Weijie et al. 2026. "Geographies of Hope: Rethinking Deepfake Harms and Gender AI Safety in the Global South." *Media and Communication* 14 (January): 10969. <https://doi.org/10.17645/mac.10969>.
- Imana, Basileal et al. 2021. *Auditing for Discrimination in Algorithms Delivering Job Ads*. Version 1. <https://doi.org/10.48550/ARXIV.2104.04502>.
- Isaksson, Anna. 2025. "Mitigation Measures for Addressing Gender Bias in Artificial Intelligence Within Healthcare Settings: A Critical Area of Sociological Inquiry." *AI & SOCIETY* 40 (4): 3009–18. <https://doi.org/10.1007/s00146-024-02067-y>.
- Iso, Hayate et al. 2025. "Evaluating Bias in LLMs for Job-Resume Matching: Gender, Race, and Education." *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, April, 672–83.
- Kalim, Usama et al. 2026. "Female Gender Bias in Artificial Intelligence Applications for Education: A Systematic Review of Regional Disparities and Equity Implications." *AI & SOCIETY* 41 (4): 2943–59. <https://doi.org/10.1007/s00146-025-02811-y>.
- Kelley, Stephanie et al. 2022. "Antidiscrimination Laws, Artificial Intelligence, and Gender Bias: A Case Study in Nonmortgage Fintech Lending." *Manufacturing & Service Operations Management* 24 (6): 3039–59. <https://doi.org/10.1287/msom.2022.1108>.
- Kelly, Sonja, and Mehrdad Mirpourian. 2021. *Algorithmic Bias, Financial Inclusion, and Gender: A Primer on Opening up New Credit to Women in Emerging Economies*. Women's World Banking.
- Kohls v. Bonta, 2:24-cv-02527-JAM-CKD (August 20, 2025). <https://docs.justia.com/cases/federal/district-courts/california/caedce/2:2024cv02527/453046/98>.
- Kong, Haein et al. 2024. "Gender Bias in LLM-generated Interview Responses." Version 3. Pre-published. <https://doi.org/10.48550/ARXIV.2410.20739>.
- Kotek, Hadas et al. 2023. "Gender Bias and Stereotypes in Large Language Models." *Proceedings of The ACM Collective Intelligence Conference* (Delft Netherlands), November 6, 12–24. <https://doi.org/10.1145/3582269.3615599>.
- Larrazabal, Agostina J. et al. 2020. "Gender Imbalance in Medical Imaging Datasets Produces Biased Classifiers for Computer-Aided Diagnosis." *Proceedings of the National Academy of Sciences* 117 (23): 12592–94. <https://doi.org/10.1073/pnas.1919012117>.
- Le, Thuy Dung, and Nicola Döring. 2026. "Perspectives on Non-consensual Deepfake Pornography: A Content Analysis of Reddit Discussions Initiated by Perpetrators, Victims, and Bystanders." *Sexuality & Culture*, ahead of print, March 4. <https://doi.org/10.1007/s12119-026-10566-x>.
- Lee, Heisook et al. 2026. "Equitable AI in Healthcare: Navigating Sex, Gender, and Intersectional Biases in Diagnostics." In *Sex, Gender, and Emerging Technology in Healthcare: Mitigating Bias and Fostering Equity*, edited by Heisook Lee et al. Springer Nature Singapore. https://doi.org/10.1007/978-981-95-2070-1_10.
- Lewicki, Kornel et al. 2023. "Out of Context: Investigating the Bias and Fairness Concerns of 'Artificial Intelligence as a Service'." *Proceedings of the 2023 CHI Conference on Human Factors in Computing*

- Systems (Hamburg Germany), April 19, 1–17. <https://doi.org/10.1145/3544548.3581463>.
- Liang, Chen et al. 2018. "Gender Wage Gap in Online Gig Economy and Gender Differences in Job Preferences." *SSRN Electronic Journal*, ahead of print. <https://doi.org/10.2139/ssrn.3266249>.
- Liu, Zilong, and Hongyan Liang. 2025. "Are Credit Scores Gender-Neutral? Evidence of Mis-Calibration from Alternative and Traditional Borrowing Data." *Journal of Behavioral and Experimental Finance* 47 (September): 101081. <https://doi.org/10.1016/j.jbef.2025.101081>.
- McInerney, Kerry. 2023. "Coding 'Carnal Knowledge' into Carceral Systems: A Feminist Abolitionist Approach to Predictive Policing." In *Feminist AI*, 1st ed., edited by Jude Browne et al. Oxford University Press. <https://doi.org/10.1093/oso/9780192889898.003.0007>.
- Mehrabi, Ninareh et al. 2022. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys* 54 (6): 1–35. <https://doi.org/10.1145/3457607>.
- Micha, Ariela et al. 2022. "When Women Enter Male-Dominated Territories in the Platform Economy: Gender Inequalities Among Drivers and Riders in Argentina." *Gender & Development* 30 (3): 575–600. <https://doi.org/10.1080/13552074.2022.2117931>.
- Microsoft, and LinkedIn. 2025. *2025 Work Trend Index Annual Report: The Year the Frontier Firm Is Born*. <https://www.microsoft.com/en-us/worklab/work-trend-index/2025-the-year-the-frontier-firm-is-born>.
- Mökander, Jakob et al. 2024. "Auditing Large Language Models: A Three-Layered Approach." *AI and Ethics* 4 (4): 1085–115. <https://doi.org/10.1007/s43681-023-00289-2>.
- Nadeem, Ayesha et al. 2022. "Gender Bias in AI-based Decision-Making Systems: A Systematic Literature Review." *Australasian Journal of Information Systems* 26 (December). <https://doi.org/10.3127/ajis.v26i0.3835>.
- Naik, Ranjita, and Besmira Nushi. 2023. "Social Biases Through the Text-to-Image Generation Lens." *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal QC Canada), August 8, 786–808. <https://doi.org/10.1145/3600211.3604711>.
- National Institute of Justice. 2021. *2021 Review and Revalidation of the First Step Act Risk Assessment Tool*. <https://www.ojp.gov/pdffiles1/nij/303859.pdf>.
- Neely, Megan Tobias et al. 2023. "Social Inequality in High Tech: How Gender, Race, and Ethnicity Structure the World's Most Powerful Industry." *Annual Review of Sociology* 49 (1): 319–38. <https://doi.org/10.1146/annurev-soc-031021-034202>.
- Noble-Tolla, Mary. 2026. "New Research: Women Use AI Less Often at Work and Get Less Credit." *Lean In*, June 5. <https://leanin.org/articles/findings/ai-women-gender-gap-data/>.
- Nondiscrimination in Health Programs and Activities (2024). <https://www.federalregister.gov/documents/2024/05/06/2024-08711/nondiscrimination-in-health-programs-and-activities>.
- OECD. 2024. *OECD Digital Economy Outlook 2024 (Volume 1): Embracing the Technology Frontier*. OECD Digital Economy Outlook. OECD Publishing. <https://doi.org/10.1787/a1689dc5-en>.
- Otokiti, Ahmed Umar et al. 2025. "Gender and Racial Bias Unveiled: Clinical Artificial Intelligence (AI) and Machine Learning (ML) Algorithms Are Fanning the Flames of Inequity." *Oxford Open Digital Health* 3 (January): oqaf027. <https://doi.org/10.1093/oodh/oqaf027>.
- Pal, Siddhi et al. 2024. *AI's Missing Link: The Gender Gap in the Talent Pool*. Data Brief. Interface. <https://www.interface-eu.org/index.php/publications/ai-gender-gap>.
- Plaza-del-Arco, Flor Miriam et al. 2024. "Angry Men, Sad Women: Large Language Models Reflect Gendered Stereotypes in Emotion Attribution." Version 3. Pre-published. <https://doi.org/10.48550/ARXIV.2403.03121>.
- Raghavan, Manish, and Solon Barocas. 2019. "Challenges for Mitigating Bias in Algorithmic Hiring." Brookings Institution, December 6. <https://www.brookings.edu/articles/challenges-for-mitigating-bias-in-algorithmic-hiring/>.
- Rashidi, Hooman H. et al. 2025. "Introduction to Artificial Intelligence and Machine Learning in Pathology and Medicine: Generative and Nongenerative Artificial Intelligence Basics." *Modern Pathology* 38 (4): 100688. <https://doi.org/10.1016/j.modpat.2024.100688>.
- Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act) (2022). <https://eur-lex.europa.eu/eli/reg/2022/2065/oj>.
- Regulation (EU) 2024/1689 of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) (2024). <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.
- Renan Barzilay, Arianne, and Anat Ben-David. 2017. "Platform Inequality: Gender in the Gig-Economy." *SSRN Electronic Journal*, ahead of print. <https://doi.org/10.2139/ssrn.2995906>.
- Richardson, Rashida et al. 2019. "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice." *New York University Law Review* 94: 15–55. <https://nyulawreview>.

[org/wp-content/uploads/2019/04/NYULawReview-94-Richardson_etal-FIN.pdf](https://www.nyulawreview.org/wp-content/uploads/2019/04/NYULawReview-94-Richardson_etal-FIN.pdf).

Rieper, Max. 2026. "How AI-generated Content Laws Are Changing Across the Country." MultiState, February. <https://www.multistate.us/insider/2026/2/12/how-ai-generated-content-laws-are-changing-across-the-country>.

Ringrose, Jessica et al. 2026. "Sexualized Deepfakes in UK Schools: Understanding and Preventing AI-Generated Image-Based Sexual Abuse Through Better AI Literacies." *Behavioral Sciences* 16 (4): 554. <https://doi.org/10.3390/bs16040554>.

Roy, Aratrika, and Priyanka Verma. 2026. "A Comprehensive Review of Structural Risk Assessment Tools for Detecting Violent Recidivism." *Journal of Forensic and Legal Medicine* 120 (May): 103162. <https://doi.org/10.1016/j.jflm.2026.103162>.

Schnepper, Rebekka et al. 2025. "Exploring Biases of Large Language Models in the Field of Mental Health: Comparative Questionnaire Study of the Effect of Gender and Sexual Orientation in Anorexia Nervosa and Bulimia Nervosa Case Vignettes." *JMIR Mental Health* 12 (March): e57986–86. <https://doi.org/10.2196/57986>.

SD, Navya et al. 2026. "Structural Gender Bias in Credit Scoring: Proxy Leakage." Version 1. Pre-published. <https://doi.org/10.48550/ARXIV.2601.18342>.

Sindoni, Maria Grazia. 2024. "The Feminization of AI-powered Voice Assistants: Personification, Anthropomorphism and Discourse Ideologies." *Discourse, Context & Media* 62 (December): 100833. <https://doi.org/10.1016/j.dcm.2024.100833>.

Sivakaminathan, Siva Sankari, and Elena Musi. 2026. "ChatGPT Is a Gender Bias Echo-Chamber in HR Recruitment: An NLP Analysis and Framework to Uncover the Language Roots of Bias." *AI & SOCIETY* 41 (4): 2841–61. <https://doi.org/10.1007/s00146-025-02564-8>.

Skeem, Jennifer et al. 2016. "Gender, Risk Assessment, and Sanctioning: The Cost of Treating Women Like Men." *Law and Human Behavior* 40 (5): 580–93. <https://doi.org/10.1037/lhb0000206>.

Sloane, Mona et al. 2022. "A Silicon Valley Love Triangle: Hiring Algorithms, Pseudo-Science, and the Quest for Auditability." *Patterns* 3 (2): 100425. <https://doi.org/10.1016/j.patter.2021.100425>.

Smith, Genevieve. 2025. "Mindsets and Management: AI and Gender (In)Equitable Access to Finance." Version 3. Pre-published. <https://doi.org/10.48550/ARXIV.2504.07312>.

Song, Ziheng et al. 2024. "Do FinTech Algorithms Reduce Gender Inequality in Banks Loans? A Quantitative Study

from the USA." *Journal of Applied Economics* 27 (1): 2324247. <https://doi.org/10.1080/15140326.2024.2324247>.

Stanczak, Karolina, and Isabelle Augenstein. 2021. "A Survey on Gender Bias in Natural Language Processing." Version 1. Pre-published. <https://doi.org/10.48550/ARXIV.2112.14168>.

Stephany, Fabian, and Jędrzej Duszynski. 2026. "Women Worry, Men Adopt: How Gendered Perceptions Shape the Use of Generative AI." Version 1. Pre-published. <https://doi.org/10.48550/ARXIV.2601.03880>.

Straw, Isabel, and Honghan Wu. 2022. "Investigating for Bias in Healthcare Algorithms: A Sex-Stratified Analysis of Supervised Machine Learning Models in Liver Disease Prediction." *BMJ Health & Care Informatics* 29 (1): e100457. <https://doi.org/10.1136/bmjhci-2021-100457>.

Sun, Luhang et al. 2023. "Smiling Women Pitching down: Auditing Representational and Presentational Gender Biases in Image-Generative AI." *Journal of Computer-Mediated Communication* 29 (1): zmad045. <https://doi.org/10.1093/jcmc/zmad045>.

Sun, Tony et al. 2019. "Mitigating Gender Bias in Natural Language Processing: Literature Review." Version 1. Pre-published. <https://doi.org/10.48550/ARXIV.1906.08976>.

Suresh, Harini, and John Guttag. 2021. "A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle." *Equity and Access in Algorithms, Mechanisms, and Optimization*, October 5, 1–9. <https://doi.org/10.1145/3465416.3483305>.

TAKE IT DOWN Act (2025). <https://www.congress.gov/119/plaws/publ12/PLAW-119publ12.pdf>.

Tang, Yanli et al. 2021. "Gender-Related Operational Issues Arising from On-Demand Ride-Hailing Platforms: Safety Concerns and System Configuration." *Production and Operations Management* 30 (10): 3481–96. <https://doi.org/10.1111/poms.13444>.

Texas Senate Bill 1188: Relating to Electronic Health Record Requirements; Authorizing a Civil Penalty (2025). <https://capitol.texas.gov/BillLookup/History.aspx?LegSess=89R&Bill=SB1188>.

The White House. 2026. *National Policy Framework for Artificial Intelligence (Legislative Recommendations)*. Washington, DC. <https://www.whitehouse.gov/wp-content/uploads/2026/03/03.20.26-National-Policy-Framework-for-Artificial-Intelligence-Legislative-Recommendations.pdf>.

Title VII of the Civil Rights Act of 1964 (1964). <https://www.eeoc.gov/statutes/title-vii-civil-rights-act-1964>.

Umbach, Rebecca et al. 2026. "AI-generated Image-Based

Sexual Abuse: Perpetration and Consumption Across Three Regions." *Computers in Human Behavior* 179 (June): 108935. <https://doi.org/10.1016/j.chb.2026.108935>.

UN Women. 2025. *How AI Is Exacerbating Technology-Facilitated Violence Against Women and Girls*. United Nations Entity for Gender Equality and the Empowerment of Women (UN Women).

UNESCO. 2022. *Recommendation on the Ethics of Artificial Intelligence*. Programme and Meeting Document SHS/BIO/PI/2021/1. United Nations Educational, Scientific and Cultural Organization. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>.

UNESCO. 2024. *Challenging Systematic Prejudices: An Investigation into Gender Bias in Large Language Models*. International Research Centre on Artificial Intelligence.

UNESCO, and EQUALS Skills Coalition. 2019. *I'd Blush If I Could: Closing Gender Divides in Digital Skills Through Education*. UNESCO. <https://doi.org/10.54675/RAPC9356>.

U.S. Department of Labor, Wage and Hour Division. 2024. *Field Assistance Bulletin No. 2024-1: Artificial Intelligence and Automated Systems in the Workplace Under the Fair Labor Standards Act and Other Federal Labor Standards*. Field Assistance Bulletin. U.S. Department of Labor. <https://www.dol.gov/agencies/whd/field-assistance-bulletins>.

U.S. Food and Drug Administration. 2013. "Drug Safety Communication: Risk of Next-Morning Impairment After Use of Insomnia Drugs; FDA Requires Lower Recommended Doses for Certain Drugs Containing Zolpidem (Ambien, Ambien CR, Edluar, and Zolpimist)." Drug Safety Communication. U.S. Food and Drug Administration, January 10. <https://www.fda.gov/files/drugs/published/Drug-Safety-Communication--Risk-of-next-morning-impairment-after-use-of-insomnia-drugs--FDA-requires-lower-recommended-doses-for-certain-drugs-containing-zolpidem-%28Ambien--Ambien-CR--Edluar--and-Zolpimist%29.pdf>.

U.S. Government Accountability Office. 2024. *Facial Recognition Technology: Federal Law Enforcement Agency Efforts Related to Civil Rights and Training*. GAO-24-107372. U.S. Government Accountability Office. <https://www.gao.gov/products/gao-24-107372>.

Varsha, P.S. 2023. "How Can We Manage Biases in Artificial Intelligence Systems – A Systematic Literature Review." *International Journal of Information Management Data Insights* 3 (1): 100165. <https://doi.org/10.1016/j.jjime.2023.100165>.

Waltz, Margaret et al. 2023. "Exclusion of Women from Phase I Trials: Perspectives from Investigators and Research Oversight Officials." *Ethics & Human Research* 45 (6): 19–30. <https://doi.org/10.1002/eahr.500170>.

Wan, Yixin et al. 2023. "'Kelly Is a Warm Person, Joseph Is a Role Model': Gender Biases in LLM-Generated Reference Letters." Version 5. Pre-published. <https://doi.org/10.48550/ARXIV.2310.09219>.

Wei, Miranda et al. 2025. "'We're Utterly Ill-Prepared to Deal with Something Like This': Teachers' Perspectives on Student Generation of Synthetic Nonconsensual Explicit Imagery." *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama Japan), April 26, 1–18. <https://doi.org/10.1145/3706598.3713226>.

Wilson, Kyra, and Aylin Caliskan. 2024. "Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (October): 1578–90. <https://doi.org/10.1609/aies.v7i1.31748>.

Wu, Gloria et al. 2025. "Chatbots and Diabetes: Is There Gender Bias?" *Journal of Patient Experience* 12 (September): 23743735251380954. <https://doi.org/10.1177/23743735251380954>.

Zack, Travis et al. 2024. "Assessing the Potential of GPT-4 to Perpetuate Racial and Gender Biases in Health Care: A Model Evaluation Study." *The Lancet Digital Health* 6 (1): e12–22. [https://doi.org/10.1016/S2589-7500\(23\)00225-X](https://doi.org/10.1016/S2589-7500(23)00225-X).

Završnik, Aleš. 2021. "Algorithmic Justice: Algorithms and Big Data in Criminal Justice Settings." *European Journal of Criminology* 18 (5): 623–42. <https://doi.org/10.1177/1477370819876762>.