POTENTIAL HARMS AND MITIGATION PRACTICES for Automated Decision-making and Generative AI

APRIL 2025

FUTURE OF PRIVACY FORUM | AI

UPDATED BY

Amber Ezzell

Policy Counsel for Artificial Intelligence, Future of Privacy Forum

---

Acknowledgments

This report benefitted from review, recommendations, and contributions from Anne J. Flanagan and Stacey Gray.

The original publication of this resource in 2017 was led by Lauren Smith at the Future of Privacy Forum.

FUTURE OF PRIVACY FORUM | CENTER FOR ARTIFICIAL INTELLIGENCE

About Future of Privacy Forum (FPF)

The **Future of Privacy Forum (FPF)** is a non-profit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. Learn more about FPF by visiting fpf.org.

About FPF's Center for Artificial Intelligence

The Center for Artificial Intelligence at the Future of Privacy Forum is dedicated to navigating the complex landscape of AI governance and its intersection with privacy and data protection law. Drawing on expertise from a global Leadership Council comprising industry leaders, academics, civil society, and policymakers, the Center provides sophisticated, practical policy analysis to help organizations align innovation with responsible implementation while meeting evolving regulatory requirements. Learn more about the FPF Center for AI at fpf.org/ai.

Overview of 2025 Update

Automated analysis of personal data, including through the use of artificial intelligence and machine learning tools, can be used to improve services, advance research, and combat discrimination. However, automated decision-making can also lead to potential harms in higher risk contexts, such as hiring, education, and healthcare, as well as differential treatment or impact on marginalized communities or vulnerable populations. When seeking to identify potential harms from development or deployment of AI tools, it is important to appreciate the context of interactions between individuals, companies, and governments — including the benefits provided by automated decision-making frameworks, and the potential fallibility of human decision-making. In other words, potential harms identified may or may not be "less harmful" than human or non-automated decision-making, and may exist across a broad spectrum of what is legal or illegal, fair or unfair, or morally or ethically responsible.

As artificial intelligence has evolved — including the rapid development and application of generative AI — new and emerging risks present complex questions regarding how to mitigate and manage such challenges. There are few easy ways to navigate these issues; however, if developers and deployers of AI tools evaluate their risk mitigation strategies to consider the vast array of potential AI harms, it can promote fairness, encourage responsible data use, and combat discrimination.

To facilitate these discussions, the Future of Privacy Forum (FPF) has updated our 2017 resource ("Distilling the Harms of Automated Decision-making") with the goal of identifying and categorizing a broad range of potential harms that may result from automated decision-making, including heightened harms related to generative AI (GenAI), and potential mitigation practices. FPF reviewed leading books, articles, and other literature on the topic of algorithmic discrimination and AI risk management, and distilled the enumerated harms and mitigation strategies identified in the literature into three tables: Potential Harms of Automated Decision-making, Heightened Risks of Generative AI, and Potential Mitigation Practices. **This resource is intended to help developers, deployers, policymakers, and the public understand the complexities of the intersection of AI and privacy, civil rights, and risk management — and to provide an accessible and comprehensive tool to help understand, design, and deploy AI tools.**

We hope you will suggest revisions, identify challenges, and help improve the document by contacting the FPF Center for Artificial Intelligence at ai@fpf.org.

Description of Tables

Potential Harms of Automated Decision-making

This table groups the harms identified in the literature into four broad "buckets" — loss of opportunity, economic loss, social detriment, and loss of liberty or life — to depict the various spheres of life where automated decision-making can cause injury. It also notes whether each harm manifests for individuals or collectives, and as illegal or unfair.

We hope that a more clear articulation of harms will help focus attention on potential mitigation strategies that can reduce the risks of algorithmic discrimination. The harms listed in this chart are not meant to be exhaustive, nor do we presume to establish which harms pose greater or lesser risks to individuals or society. These harms can — and often do — occur in tandem with one another.

Heightened Risks of Generative AI

In conjunction with the Table of Potential Harms of Automated Decision-making, this table highlights some of the additional risks associated with Generative AI.

The full range of individual and societal risks and benefits of generative AI are still being understood. The table includes a description of some of the challenges and ethical questions presented by generative AI tools. Because law and policy is still evolving in this area, these harms are ripe for further analysis.

Potential Mitigation Practices

This table provides a sampling of mitigation strategies based on the relevant stage in the AI lifecycle: training and development; deployment; and post-deployment and evaluation.

The mitigation strategies provided are intended to address a range of harms. Nevertheless, the mitigation practices provided are not meant to be an exhaustive list for each respective stage in the AI development and deployment lifecycle, but rather, a sampling of responsible AI risk management practices. *The mitigation practices legally required may vary by jurisdiction.*

Potential Harms of Automated Decision-making

INDIVIDUAL HARMS		COLLECTIVE/ SOCIETAL HARMS
Illegal/Unlawful Represents actions that are illegal under several civil rights laws, which generally protect core classifications — such as race, gender, age, and ability — against discrimination, disparate treatment, and disparate impact.	**Unfair** Represents actions that are typically legal, but nonetheless trigger notions of unfairness. Like the "illegal" category, some examples here may be differently classified depending on the legal regime.	
LOSS OF OPPORTUNITY		
Employment Discrimination		**Differential Access to Job Opportunities**
E.g. Use of a hiring tool that automatically excludes job candidates based on birth year	E.g. Filtering candidates by work proximity leads to excluding people of color	
Insurance & Social Benefit Determination		**Differential Access to Insurance & Benefits**
E.g. Algorithm sets higher premiums or rates for people of color	E.g. Increasing auto insurance prices for night-shift workers	
Housing Discrimination		**Differential Access to Housing**
E.g. Landlord relies on tool that makes housing determination made in whole or in part based on protected characteristic	E.g. Matching algorithm less likely to provide suitable housing for marginalized communities	
Education Discrimination		**Differential Access to Education**
E.g. Denial of opportunity for a student in a certain ability category	E.g. Presenting only ads for for-profit colleges to low-income individuals	
ECONOMIC LOSS		
Credit Discrimination		**Differential Access to Credit**
E.g. Denying credit to all residents in specified neighborhoods ("redlining")	E.g. Not presenting certain credit offers to members of certain groups, or unfairly referencing others	
Differential Pricing of Goods and Services		**Differential Access to Goods and Services**
E.g. Raising online prices based on membership in a protected class	E.g. Presenting product discounts based on "ethnic affinity"	
	Narrowing of Choice E.g. Algorithms that prevent users from discovering new products outside their established patterns	**Narrowing of Choice for Groups**
SOCIAL DETRIMENT		
	Network Bubbles E.g. Varied exposure to opportunity or evaluation based on "who you know"	**Filter Bubbles** E.g. Algorithms that promote only familiar news and information
	Dignitary Harms E.g. Emotional distress due to bias or a decision based on incorrect data	**Stereotype Reinforcement** E.g. Assumption that computed decisions are inherently less biased than human decisions
	Constraints of Bias E.g. Overly constrained conceptions of career prospects based on early-childhood educational surveys	**Confirmation Bias** E.g. Generative AI tool produces all-male images for "CEO," all-female results for "teacher"
LOSS OF LIBERTY OR LIFE		
	Constraints of Suspicion E.g. Emotional, dignitary, and social impacts of increased surveillance	**Increased Surveillance** E.g. Use of "predictive policing" to police minority neighborhoods more
Individual Incarceration E.g. Use of "recidivism scores" to determine prison sentence length (legal status uncertain)		**Disproportionate Incarceration** E.g. Incarceration of groups at higher rates based on historic policing data
Healthcare Discrimination		**Differential Access to Healthcare**
E.g. Medical recommendations made for a patient based solely on race, color, national origin, or insurance status	E.g. Algorithm used to detect skin cancer less accurate for patients with darker skin	
Physical Safety		**Differential Access to Safety** E.g. Routing of emergency services based on optimizing route efficiency
	E.g. Missed classification/detection of pedestrians or vehicles due to skin color	

Heightened Risks of Generative AI

Generative AI, or AI systems designed to create new content, ideas, or data (e.g. text, images, or videos), can exacerbate existing risks when used in the context of automated decision-making. For example, a generative AI tool could be used to summarize patient records, recommend a personalized treatment plan, or write police reports. Such uses may present similar risks of bias and discrimination, as well as implicate additional risks related to privacy, accuracy, and safety. The full range of individual and societal risks of generative AI, as well as the potential benefits, are still being understood, and this list does ***not*** include many of the growing debates around existential risk or alignment with human values.

MISINFORMATION

AI-generated text, images, videos, or voices can be used to generate or spread false information, impacting trust in individuals, media, and institutions.

ACCURACY

Particularly when used in sensitive settings (e.g., a health treatment plan or personalized education plan), generative AI raises key accuracy challenges related to, e.g., hallucination, logical errors, misunderstandings, inconsistency, and training data biases.

SAFETY

While generative AI tools can be used to promote safety, they can also be used to further physically and mentally unsafe conduct (e.g. a chatbot used for mental health treatment providing inappropriate responses; a generative AI tool being prompted to provide instructions for building weapons).

FRAUD & SCAMS

AI-generated content can be used to impersonate individuals or organizations in order to defraud consumers; for example, AI impersonations can convince individuals to share financial information.

PRIVACY & SECURITY

Generative AI raises novel privacy and security vulnerabilities for training, fine-tuning, and for the ability to cause disclosure of personal information through extraction from a model's output.

INTELLECTUAL PROPERTY

Generative AI raises novel questions related to intellectual property (copyright, trademarks, and patents), including the lawfulness of training on protected content, and the ability for users to generate outputs implicating intellectual property law.

Potential Mitigation Practices

The following mitigation practices are not meant to be an exhaustive list for each respective stage in the AI development and deployment lifecycle, but rather, a sampling of responsible AI risk management practices that can be employed as appropriate in the context of particular AI systems and risks. ***The mitigation practices legally required, where applicable, may vary by jurisdiction.***

TRAINING & DEVELOPMENT

- Algorithmic design with informed human oversight and engagement **("human in the loop")** to enhance the explainability, transparency, and accountability
- **Clearly defined responsibilities** for developers to ensure proper management and oversight of AI tools
- **Data methods** to ensure proxies are not used for protected classes, and training data does not amplify historical bias
- **Human rights impact assessments** to assess the impact on fundamental rights that the system may produce
- **Privacy impact assessments** to ensure that personal data is collected, used, shared, and maintained solely within the scope of organizational privacy policy and consistent with jurisdictional requirements
- **Use of DPIAs** to measure impact or enable rights to explanation
- **Transparency** to provide deployers — and when feasible, impacted individuals — with information about how the tool is fit for purpose, addresses bias, calculates risk, and attempts to limit harm
- **Red teaming** and other testing to protect against a range of harms, and to promote privacy and security, making adjustments to tools based on findings
- **Audits** to assess whether input data results in bias, disparate treatment, or disparate impact of certain protected groups
- **Conformity assessments** to determine whether the AI system meets legal and ethical standards prior to being placed on the market
- Designing AI tools with **alternative review procedures** in mind

DEPLOYMENT

- Algorithmic operation with informed human oversight and engagement **("human in the loop")** to enhance the explainability, transparency, and accountability
- **Clearly defined responsibilities** for deployers to ensure proper management and oversight of AI tools
- **Testing at various stages of deployment** to ensure tools are fit for purpose and assessed for disparate treatment and impact (e.g. race, gender, sexual orientation, gender identity, disability, age, religion, socioeconomic status, national origin, etc.)
- **Human rights impact assessments** to assess the impact on fundamental rights that the system may produce
- **Privacy impact assessments** to ensure that personal data is collected, used, shared, and maintained solely within the scope of organizational privacy policy and consistent with jurisdictional requirements
- **Disclosure to affected entity (e.g. vendor, user, etc.)** when AI system is found to result in disparate impact or disparate treatment
- **Audits** to assess whether the use of a tool results in bias, disparate treatment, or disparate impact of certain protected groups
- Configuring AI tools with **alternative review procedures** for individuals who legally require reasonable accommodations

POST-DEPLOYMENT & EVALUATION

- **Internal business processes** to index concerns; ethical frameworks & best practices to monitor and evaluate outcomes
- **Regular review of high-risk systems** to ensure the system does not engage in algorithmic discrimination
- **Privacy impact assessments** to ensure that personal data is collected, used, shared, and maintained solely within the scope of organizational privacy policy and consistent with jurisdictional requirements
- **Human rights impact assessments** to assess the impact on fundamental rights that the system may produce
- **Audits** to assess whether the use of a tool results in bias, disparate treatment, or disparate impact of certain protected groups
- **Transparency to developers, deployers, and/or consumers** about the results of bias audits performed

Working Definitions

The following definitions reflect how these terms are used within this resource.

Automated Decision

The output or indirect result of a machine-based system that makes predictions, recommendations, or decisions for a given set of objectives without the direct involvement of a human being.

Illegal

Examples in this category represent harms that are generally illegal under several civil rights laws, which generally protect core classifications — such as race, gender, age, and ability — against discrimination, disparate treatment, and disparate impact. Classification as illegal versus unfair may vary based on the legal regime.

Unfair

Examples in this category represent actions that are typically legal, but nonetheless trigger notions of unfairness. Classification as unfair versus illegal may vary based on the legal regime.

Collective/Societal Harms

This category represents overall negative effects to society that are chiefly collective, rather than individual in nature.

Loss of Opportunity

This group broadly describes harms occurring within domains such as the workplace, housing, social support systems, healthcare, and education.

Economic Loss

This group broadly describes harms that primarily cause financial injury or discrimination in the marketplace for goods and services.

Social Detriment

This group broadly describes harms that impact one's sense of self, self worth, well-being, or community standing relative to others.

Loss of Liberty or Life

This group broadly describes harms that constrain one's physical freedom, autonomy, or may pose a threat to human life.

Reviewed Literature

The alphabetized list below captures the literature FPF has reviewed to date for this effort. We welcome suggestions for further materials to review to the FPF Center for Artificial Intelligence at ai@fpf.org.

Literature Reviewed for 2025 Update

- "AI and the Risk of Consumer Harm," Federal Trade Commission (January 3, 2025), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2025/01/ai-risk-consumer-harm
- "Background Dossiers and Algorithmic Scores for Hiring, Promotion, and Other Employment Decisions" (Consumer Financial Protection Circular 2024-06), https://www.consumerfinance.gov/compliance/circulars/consumer-financial-protection-circular-2024-06-background-dossiers-and-algorithmic-scores-for-hiring-promotion-and-other-employment-decisions/
- Chiraag Bains, Brookings Institution, The legal doctrine that will be key to preventing AI discrimination (Sept. 13, 2024), https://www.brookings.edu/articles/the-legal-doctrine-that-will-be-key-to-preventing-ai-discrimination
- Danielle Keats Citron and Daniel J. Solove, Privacy Harms (February 9, 2021). GWU Legal Studies Research Paper No. 2021-11, GWU Law School Public Law Research Paper No. 2021-11, 102 Boston University Law Review 793 (2022), https://ssrn.com/abstract=3782222
- EPIC, Generating Harms: Generative AI's Impact & Paths Forward, May 2023, https://epic.org/wp-content/uploads/2023/05/EPIC-Generative-AI-White-Paper-May2023.pdf
- Equal Employment Opportunity Commission, "The Americans with Disabilities Act and the Use of Software, Algorithms, and Artificial Intelligence to Assess Job Applicants and Employees," May 12, 2022, https://www.eeoc.gov/laws/guidance/americans-disabilities-act-and-use-software-algorithms-and-artificial-intelligence
- Future of Privacy Forum, Best Practices for AI and Workplace Assessment Technologies, https://fpf.org/wp-content/uploads/2024/02/FPF-Best-Practices-for-AI-and-WA-Tech-FINAL-with-date.pdf, November 2023.
- Gil Appel, Juliana Neelbauer, and David A. Schweidel, "Generative AI Has an Intellectual Property Problem," Harvard Business Review (Apr. 7, 2023), https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem
- Guidance on Application of the Fair Housing Act to the Advertising of Housing, Credit, and Other Real Estate-Related Transactions through Digital Platforms," U.S. Department of Housing and Urban Development (April 29, 2024), https://www.hud.gov/sites/dfiles/FHEO/documents/FHEO_Guidance_on_Advertising_through_Digital_Platforms.pdf
- "Guidance on Application of the Fair Housing Act to Screening of Applicants for Rental Housing," U.S. Department of Housing and Urban Development (April 29, 2024), https://www.hud.gov/sites/dfiles/FHEO/documents/FHEO_Guidance_on_Screening_of_Applicants_for_Rental_Housing.pdf, 5
- Joint Statement on Enforcement of Civil Rights, Fair Competition, Consumer Protection, and Equal Opportunity Laws in Automated Systems, April 4, 2024, https://www.justice.gov/crt/media/1346821/dl?inline
- MIT AI RIsk Repository, https://airisk.mit.edu/
- "Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964," U.S. Equal Employment Opportunity Commission (May 18, 2023), https://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-algorithms-and-artificial
- U.S. Department of Education Office of Educational Technology, "Artificial Intelligence and the Future of Teaching and Learning: Insights and Recommendations," May 2023, https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf

- UK Department for Science, Innovation & Technology, "Guidance: Introduction to AI assurance," (Feb. 12, 2024), https://www.gov.uk/government/publications/introduction-to-ai-assurance/introduction-to-ai-assurance
- U.S. Department of Education, Office for Civil Rights, "Avoiding the Discriminatory Use of Artificial Intelligence," November 2024, https://www.ed.gov/media/document/avoiding-discriminatory-use-of-ai
- U.S. Department of Labor, Partnership on Employment & Accessible Technology (PEAT), AI & Inclusive Hiring Framework, https://www.peatworks.org/ai-inclusive-hiring-framework/framework-overview/
- "What is automated individual decision-making and profiling?," Information Commissioner's Office, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/automated-decision-making-and-profiling/what-is-automated-individual-decision-making-and-profiling/
- "What is the EEOC's role in AI?," U.S. Equal Employment Opportunity Commission (April 29, 2024), https://www.eeoc.gov/sites/default/files/2024-04/20240429_What%20is%20the%20EEOCs%20role%20in%20AI.pdf

Literature Reviewed for 2017 Publication

- Aaron Reike, Don't let the hype over "social media scores" distract you, EQUAL FUTURE (2016)
- Alessandro Acquisti & Christina Fong, An Experiment in Hiring Discrimination via Online Social Network, presented at Privacy Law Scholars Conference (2016)
- Alethea Lange et al., A User-Centered Perspective on Algorithmic Personalization, presented at the Fed. Trade Comm'n PrivacyCon Conference (2017)
- Allan King & Marko Mrkonich, "Big Data" and the Risk of Employment Discrimination, 68 OKLA. L. REV. 555 (2016)
- Andrew Tutt, An FDA for Algorithms, 67 ADMIN. L. REV. 1 (2016)
- Aniko Hannak et al., Bias in Online Freelance Marketplaces: Evidence from TaskRabbit, presented at the Workshop on Data and Algorithmic Transparency (Nov. 2016)
- Cathy O'Neil, Weapons of Math Destruction (2016)
- Christian Sandvig et al., Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms, presented at the Int'l Comm'cn Ass'n Conference on Data and Discrimination: Converting Critical Concerns into Productive Inquiry (2014)
- Daniel Solove, A Taxonomy of Privacy, 154 U. PENN. L. REV. 3 (2016)
- Danielle Keats Citron & Frank Pasquale, The Scored Society: Due Process for Automated Predictions, 89 WASH. L. REV. 1 (2014)
- EXEC. OFF. OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES (2014). • EXEC. OFF. OF THE PRESIDENT, BIG DATA: A REPORT ON ALGORITHMIC SYSTEMS, OPPORTUNITY, AND CIVIL RIGHTS (2016)
- FEDERAL TRADE COMMISSION, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? (Jan. 2016)
- Frank Pasquale & Danielle Keats Citron, Promoting Innovation While Preventing Discrimination: Policy Goals for the Scored Society, 89 WASH. L. REV. 1413 (2014)
- Jennifer Valentino-Devries, Jeremy Singer-Vine, Ashkan Soltani, Websites Vary Prices, Deals Based on Users' Information, WALL ST. J. (Dec. 24, 2012)
- Joshua Kroll et al., Accountable Algorithms, 165 U. PENN. L. REV. 633 (2016)
- Juhi Kulshrestha et al., Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media, presented at the Workshop on Data and Algorithmic Transparency (2016)

- Kate Crawford & Jason Schultz, Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms, 55 B.C.L. REV. 93 (2014)
- Latanya Sweeney, Discrimination in Online Ad Delivery, COMMC'NS OF THE ASS'N OF COMPUTING MACHINERY (2013)
- Lee Rainie & Jana Anderson, Code-Dependent: Pros and Cons of the Algorithm Age, PEW RESEARCH CENTER (2017)
- Mark MacCarthy, Student Privacy: Harm and Context, 21 INT'L REV. OF INFO. ETHICS 11 (2014)
- Mary Madden, Michele Gilman, Karen Levy & Alice Marwick, Privacy , Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans, Wash. U. L. Rev __ (forthcoming) (Mar. 2017)
- Megan Garcia, How to Keep Your AI From Turning Into a Racist Monster, WIRED (2017)
- Moritz Hardt, Eric Price & Nathan Srebro, Equality of Opportunity in Supervised Learning, presented at the Conference on Neural Info. Processing Sys. (2016)
- Motahhare Eslami et al., Reasoning about Invisible Algorithms in the News Feed, presented at the Ass'n of Computing Machinery Special Interest Gp. on Computer-Human Interaction (2015)
- Muhammad Zafar et al., Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment, presented at the Int'l World Wide Web Conference (2017)
- Nanette Byrnes, Why We Should Expect Algorithms to be Biased, MIT TECHNOLOGY REVIEW (2016) • NEW AMERICA & OPEN TECH. INST., DATA AND DISCRIMINATION: COLLECTED ESSAYS (S.P. Gangadharan, Ed. 2014)
- Omer Tene and Jules Polonetsky, Big Data for All: Privacy and User Control in the Age of Analytics, 11 Nw. J. Tech. & Intell. Prop.239 (2013)
- PAM DIXON & ROBERT GELLMAN, THE SCORING OF AMERICA: HOW SECRET CONSUMER SCORES THREATEN YOUR PRIVACY AND YOUR FUTURE, WORLD PRIVACY FORUM (2014)
- Pauline Kim, Data-Driven Discrimination at Work, 59 WILLIAM & MARY L. REV. ___ (2017)
- Peter Swire, Lessons From Fair Lending Law for Fair Marketing and Big Data (2014)
- PROPUBLICA, Machine Bias Investigative Series, https://www.propublica.org/series/machine-bias
- Sandra Wachter, Brent Mittelstadt, & Luciano Floridi, Why a right to explanation of automated decision making does not exist in the General Data Protection Regulation (2016)
- Solon Barocas & Andrew Selbst, Big Data's Disparate Impact, 104 CALIF. L. REV. 671 (2016)