FUTURE OF PRIVACY FORUM | AI

Potential Harms of Automated Decision-making

INDIVIDUAL 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.

COLLECTIVE/ SOCIETAL HARMS

LOSS OF OPPORTUNITY

Illegal/Unlawful	Unfair	Collective/Societal Harms
Employment Discrimination 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	**Differential Access to Job Opportunities**
Insurance & Social Benefit Determination E.g. Algorithm sets higher premiums or rates for people of color	E.g. Increasing auto insurance prices for night-shift workers	**Differential Access to Insurance & Benefits**
Housing Discrimination 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	**Differential Access to Housing**
Education Discrimination 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	**Differential Access to Education**

ECONOMIC LOSS

Credit Discrimination 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 Access to Credit**
Differential Pricing of Goods and Services E.g. Raising online prices based on membership in a protected class	E.g. Presenting product discounts based on "ethnic affinity"	**Differential Access to Goods and Services**
	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 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	**Differential Access to Healthcare**
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	

FUTURE OF PRIVACY FORUM | AI

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