

Taxonomy of AI Harms

Type: Classification Framework

Target Audience: Risk Managers, AI Ethics Teams, Policy Makers

This framework provides a comprehensive classification system for understanding and categorizing the different types of harms that can result from AI systems. Use this taxonomy to ensure complete coverage when assessing AI risks and developing mitigation strategies.

1. Framework Overview

AI harms can be classified across multiple dimensions: by who is affected, how the harm occurs, when in the AI lifecycle it manifests, and its reversibility. This multi-dimensional approach ensures comprehensive risk identification.

Dimension	Categories	Purpose
By Affected Party	Individual, Group, Societal, Environmental	Identifies who bears the impact
By Mechanism	Allocative, Representational, Quality-of-Service	Explains how the harm occurs
By Lifecycle Stage	Design, Training, Deployment, Use	Identifies when harm is introduced
By Reversibility	Reversible, Partially Reversible, Irreversible	Assesses ability to remedy

2. Harms by Mechanism

Understanding how harm occurs helps identify appropriate mitigations.

Allocative Harms

Unfair distribution of resources, opportunities, or information based on AI decisions.

Harm Type	Description	Examples
Economic Denial	Denial of financial opportunities or resources	Credit denial, insurance pricing discrimination, loan rejection
Opportunity Denial	Blocking access to jobs, education, or services	Resume screening bias, admissions filtering, housing denial
Information Asymmetry	Unequal access to information or recommendations	Biased search results, filtered news, targeted pricing

Representational Harms

Harms to dignity, identity, or how groups are perceived and portrayed.

Harm Type	Description	Examples
Stereotyping	Reinforcing negative or limiting stereotypes	Biased image generation, gendered job suggestions
Denigration	Producing content that demeans or insults groups	Offensive outputs, derogatory associations
Erasure	Excluding or making groups invisible	Missing representation, failure to recognize
Misrepresentation	Inaccurate or distorted portrayal of groups	Cultural appropriation, historical inaccuracies

Quality-of-Service Harms

Unequal performance or reliability of AI systems across different users or contexts.

Harm Type	Description	Examples
Performance Disparity	System works better for some groups than others	Voice recognition fails for accents, facial recognition errors
Accessibility Failure	System excludes users with disabilities or limitations	No screen reader support, complex interfaces
Language Disparity	Reduced quality for non-dominant languages	Poor translation, limited language support

3. Harm Severity Classification

Classify identified harms by severity to prioritize mitigation efforts.

Level	Criteria	Examples	Response
Critical	Threat to life, safety, or fundamental rights	Medical misdiagnosis, wrongful arrest, physical injury	Immediate halt; human review required
High	Significant financial or reputational damage	Credit denial, job rejection, public defamation	Urgent remediation; escalate to leadership
Medium	Meaningful negative impact, recoverable	Poor service quality, minor financial loss	Planned remediation; monitor closely
Low	Minor inconvenience, easily corrected	Incorrect recommendation, minor error	Document and address in regular cycle

4. Harm Assessment Checklist

Use this checklist when evaluating AI systems for potential harms.

Allocative Harm Assessment

- Does the system make decisions about resource allocation?
- Could decisions systematically disadvantage protected groups?
- Are there appeals or override mechanisms for adverse decisions?
- Is there human review for high-stakes decisions?

Representational Harm Assessment

- Does the system generate or display content about people or groups?
- Have outputs been tested for stereotyping or bias?
- Are underrepresented groups adequately included in training data?
- Is there a process to identify and remove denigrating content?

Quality-of-Service Assessment

- Has performance been tested across different user demographics?
- Are there accessibility accommodations for users with disabilities?

- Does the system perform equitably across languages and accents?
- Are performance disparities monitored in production?

Assessment Completed By: _____ **Date:** _____

Reviewed By: _____ **Date:** _____