Foundations of AI Ethics & Society

An examination of the ethical, social, and institutional dynamics shaping the development, deployment, and impact of artificial intelligence.

Ghassem Tofighi

Contents

Overview

This lesson examines the interplay between artificial intelligence systems and human society. It covers ethical considerations in AI development and deployment, societal opportunities and challenges arising from AI, sources and consequences of algorithmic bias, impacts on marginalized groups, mitigation strategies, the sociotechnical nature of AI systems, basic ethical frameworks, and governance tools such as audits and impact assessments.

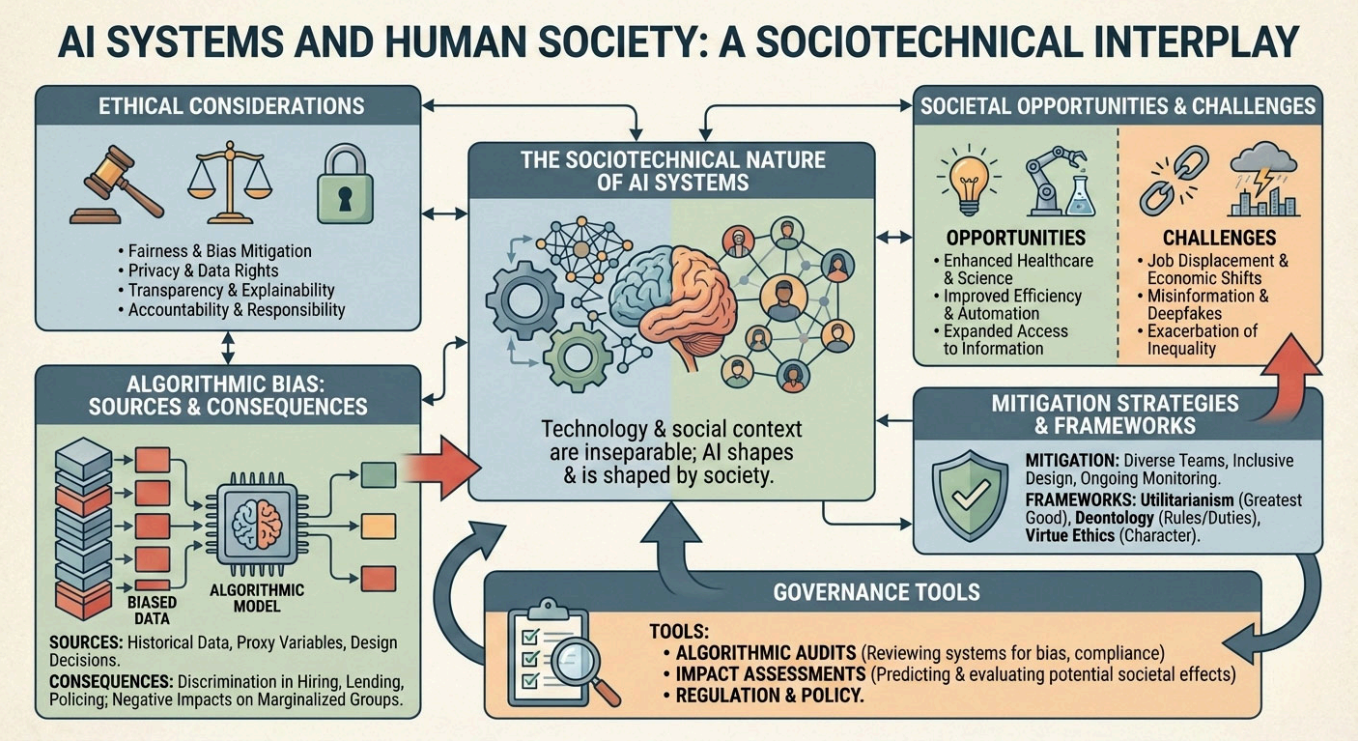

Figure 1: The Sociotechnical Interplay of AI and Society

Figure 1 details the intricate relationship between artificial intelligence and the human social structures it inhabits. Central to the diagram is the sociotechnical nature of AI, which suggests that technical systems cannot be understood in isolation from their social context. The flow begins with the sources of algorithmic bias, such as flawed historical data, and tracks how these lead to unequal impacts on marginalized groups. To address these risks, the graphic identifies key ethical frameworks like deontology and utilitarianism alongside practical governance tools. These tools, including algorithmic audits and impact assessments, serve as essential checks to maximize societal opportunities while minimizing challenges like job displacement and misinformation. By viewing AI as a cycle of mutual influence, the visual emphasizes the need for continuous monitoring and inclusive design to foster a fair technological future.

Learning Objectives

- Identify key ethical considerations that arise during the development and deployment of AI systems.
- Describe major societal opportunities and challenges created by widespread AI adoption.
- Explain the primary sources of bias in AI algorithms and provide concrete examples.
- Analyze how AI bias affects marginalized groups and society as a whole.
- Apply strategies to detect, measure, and mitigate bias and other forms of harm in AI systems.

- Define sociotechnical systems and explain how technology interacts with people, institutions, norms, and infrastructure.
- Use simple ethical frameworks to evaluate AI applications in everyday contexts.
- Describe governance tools, including algorithmic audits and algorithmic impact assessments.

Motivation

AI systems increasingly influence decisions in hiring, lending, criminal justice, healthcare, education, and public services. These systems shape access to opportunities and resources. When poorly designed or deployed, they can amplify existing inequalities or create new forms of harm. Understanding ethical dimensions and sociotechnical dynamics enables the development of AI that supports societal well-being rather than undermining it.

Ethical Considerations in AI Development and Deployment

Ethical considerations arise at every stage of the AI lifecycle. These considerations shape the design, implementation, and long-term operation of AI systems.

Data Collection and Use

Data forms the foundation of most AI systems. Ethical issues include:

- **Informed consent**: Individuals whose data is collected must understand how their information will be used and have the ability to provide or withhold consent.
- **Privacy protection**: Sensitive personal information must be handled with appropriate safeguards, including anonymization, differential privacy techniques, and secure storage.
- **Representation in training data**: Datasets must include adequate representation of all relevant demographic groups to avoid systematic exclusion or misrepresentation.
- **Data provenance and legitimacy**: Data must be obtained through lawful means without violating terms of service, scraping restrictions, or intellectual property rights.

Model Design

Choices made during model development carry ethical weight:

- **Objective function selection**: The loss function or reward signal defines what the model optimizes. Objectives that ignore fairness or equity can produce discriminatory outcomes even with balanced data.
  Example: A resume screening model optimized solely to maximize prediction of hiring success may favor candidates from historically advantaged groups if past hiring data reflects systemic bias, even when training data contains equal numbers of applicants from all groups.
- **Feature selection and engineering**: Features that serve as proxies for protected attributes can lead to indirect discrimination.
  Example: Using zip code as a feature in a credit scoring model can indirectly discriminate on the basis of race or ethnicity, because certain postal codes are strongly correlated with racial demographics due to historical redlining practices.
- **Trade-offs between performance and fairness**: Higher overall accuracy may come at the cost of worse performance on minority groups.
  Example: A medical diagnostic model may achieve 95% overall accuracy by performing very well on

the majority group (e.g., 97% for White patients) but poorly on underrepresented groups (e.g., 75% for Black patients), creating disparate error rates across populations.

- **Model complexity and interpretability**: Black-box models make it difficult to understand decision rationales, complicating accountability.
  Example: Deep neural networks used in parole risk assessment may produce accurate predictions overall but provide no clear explanation for why a particular individual is classified as high-risk, making it impossible for judges or affected persons to challenge or understand the reasoning.

Deployment Decisions

The context in which a model is used determines its real-world impact:

- **Appropriate use context**: Models trained for one purpose may produce harmful outcomes when applied in different domains.
  Example: A facial recognition model developed for security at public events may be repurposed for identifying students in school hallways, leading to disproportionate surveillance of minority students and violations of privacy expectations.
- **Affected populations**: Systems must be evaluated for differential impact across demographic groups, geographic regions, and socioeconomic categories.
  Example: An automated job recommendation system deployed nationwide may disadvantage rural applicants because it was trained primarily on urban job data and fails to recognize valid qualifications common in rural economies.
- **Human oversight mechanisms**: Decisions with high-stakes consequences require human review processes.
  Example: In automated welfare eligibility systems, final benefit denial decisions should include mandatory human review when the system flags an application as suspicious, to prevent erroneous terminations of critical support for vulnerable individuals.
- **Transparency to end-users**: Individuals subject to AI decisions should be informed that an AI system is involved and understand its limitations.
  Example: When a bank uses an AI model to approve or deny loan applications, applicants must be notified that an automated system made the decision and should be provided with information about how to request a human review or explanation.

Ongoing Maintenance

AI systems require continuous ethical attention after deployment:

- **Concept and data drift detection**: Changes in underlying data distributions or societal norms can degrade model performance and fairness over time.
  Example: A hiring model trained on pre-pandemic resume data may become unfair after widespread adoption of remote work, as it penalizes candidates who lack recent in-office experience even though remote work has become the norm.
- **Periodic re-evaluation and updating**: Models must be retrained or adjusted to maintain acceptable performance and fairness levels.
  Example: A predictive policing model should be systematically re-evaluated every six months using

current arrest and crime data to detect whether it continues to disproportionately target certain neighborhoods as societal patterns change.

- **Decommissioning criteria**: Clear conditions must be established for when a system should be retired, especially when it no longer meets ethical standards.
  Example: A facial analysis system for emotion detection in job interviews should be decommissioned if independent audits consistently show it performs poorly across racial groups and cannot be improved to acceptable levels.
- **Incident response processes**: Mechanisms must exist to detect, investigate, and remediate harms when they occur.
  Example: When users report that a content moderation AI is systematically removing posts from certain political viewpoints, the organization must have a documented process to investigate the issue, measure the extent of harm, and implement corrective actions within a defined timeframe.

Key Ethical Principles

Several principles are commonly referenced in AI ethics guidelines:

- **Fairness**: AI systems should avoid unjustified differential treatment across protected groups.
  Example: A hiring algorithm must not systematically disadvantage candidates based on gender or race, even when overall performance metrics appear strong. For instance, if the system selects male applicants at a higher rate than equally qualified female applicants for technical roles, it violates fairness despite high predictive accuracy.
- **Accountability**: Organizations and individuals developing and deploying AI must be answerable for the system's outcomes.
  Example: When an autonomous vehicle causes an accident, the company that developed the AI system, as well as the engineers who designed its decision-making algorithms, must be able to explain the system's behavior and accept responsibility for harm caused, rather than attributing fault solely to the technology itself.
- **Transparency**: The functioning, capabilities, and limitations of AI systems should be understandable to relevant stakeholders.
  Example: In a credit scoring system, applicants who are denied loans should receive a clear explanation of the key factors that influenced the decision (e.g., debt-to-income ratio, payment history) rather than receiving a vague statement that an AI model made the determination.
- **Privacy**: Personal data should be collected, used, and stored in ways that respect individual privacy rights.
  Example: A health recommendation app that uses users' location data to suggest nearby gyms must obtain explicit consent for tracking location, limit data retention periods, and provide options to delete data, rather than continuously collecting and selling location histories without user knowledge.
- **Robustness**: Systems should perform reliably under a range of conditions, including adversarial inputs and distributional shifts.
  Example: A facial recognition system used at airport security must maintain high accuracy even when individuals wear glasses, masks, or different hairstyles, and should resist adversarial attacks where small, imperceptible changes to input images cause misclassification.

- **Beneficence and non-maleficence**: AI should be developed to maximize benefits and minimize harm to individuals and society.
  Example: An AI system for mental health support must be designed to avoid providing harmful advice (e.g., suggesting dangerous coping strategies to users expressing suicidal thoughts) and should include escalation mechanisms to connect users with professional human help when a serious risk is detected.

Additional Examples of Ethical Challenges

- **Large language models**: Training on vast internet text can reproduce harmful stereotypes, toxic language, or copyrighted material without proper attribution.
- **Generative AI for images**: Systems can produce deepfakes or culturally insensitive content when prompts include protected characteristics.
- **Automated decision systems in public services**: Welfare allocation algorithms have denied benefits to eligible individuals due to technical errors or biased training data.
- **Autonomous weapons systems**: The delegation of lethal force decisions to AI raises questions of moral responsibility and meaningful human control.
- **AI in mental health applications**: Chatbots providing therapy-like interactions may lack safeguards against giving harmful advice to vulnerable users.

Societal Impacts of AI: Opportunities and Challenges

Opportunities

- Increased efficiency in healthcare diagnostics, transportation, agriculture, and manufacturing
- Expanded access to education and information through personalized learning and translation
- Enhanced scientific discovery through pattern recognition in large datasets
- Improved disaster response and resource allocation

Challenges

- Displacement of workers in routine cognitive and physical tasks
- Concentration of economic power in few organizations controlling AI infrastructure
- Erosion of human autonomy through pervasive surveillance and behavioral prediction
- Amplification of misinformation and polarization
- Environmental costs from training large models

Bias in AI Algorithms: Sources and Examples

Sources of Bias

1. **Data bias**
   - Underrepresentation or misrepresentation of certain groups
   - Historical patterns reflecting societal inequalities
   - Proxy variables correlating with protected attributes
2. **Modeling bias**
   - Choice of features that encode protected attributes
   - Optimization objectives that ignore fairness
   - Model architectures that amplify certain patterns
3. **Human-in-the-loop bias**

   - Labeler bias in supervised learning
   - Feedback loops in interactive systems
4. **Deployment bias**
   - Different error costs across groups
   - Use in contexts different from training distribution

Examples

- Facial recognition systems with higher error rates for darker skin tones and women
- Credit scoring models denying loans based on zip codes correlated with race
- Hiring algorithms favor candidates with word patterns associated with male applicants
- Predictive policing tools targeting neighborhoods based on historical arrest data

Impact of AI Bias on Marginalized Groups and Society

Biased AI systems can:

- Deny opportunities (employment, housing, education, credit)
- Increase surveillance and policing in specific communities
- Reinforce stereotypes in media and content recommendation
- Reduce trust in institutions relying on AI decisions
- Exacerbate existing social inequalities

Broader societal effects include:

- Erosion of democratic processes through targeted manipulation
- Decreased social cohesion
- Widening economic gaps between those who control AI and those affected by it

Strategies to Mitigate AI Bias and Harm

1. **Pre-processing approaches**
   These methods modify the training data before model training to reduce bias.
   - **Re-weighting or re-sampling datasets**: Adjust the sampling probability or assign higher weights to underrepresented groups to balance their influence during training.
     Example: In a loan approval dataset where women represent only 30% of accepted applications due to historical bias, re-weighting increases the contribution of female applicants during training so the model learns patterns from them more strongly.
   - **Removing or transforming protected attributes**: Suppress or transform features that directly or indirectly encode protected characteristics.
     Example: In a hiring model, remove gender-specific pronouns from resume text and replace them with neutral terms, or apply massaging techniques to make salary history distributions similar across gender groups.
2. **In-processing approaches**
   These methods incorporate fairness directly into the model training process.
   - **Fairness constraints during optimization**: Add constraints or regularization terms to the objective function that enforce fairness metrics.
     Example: During logistic regression training for credit scoring, add a constraint that the true positive

rate must be equal across racial groups (equal opportunity constraint), and optimize under this restriction using Lagrangian methods.

- **Adversarial debiasing**: Train a model to predict the target while simultaneously training an adversary to fail at predicting protected attributes from the model's representations.
  Example: In a facial analysis model, an encoder produces representations that the main classifier uses to predict age, while an adversary tries (and fails) to predict race from the same representations, forcing the encoder to remove race-related information.

3. **Post-processing approaches**
   These methods adjust model outputs after training without modifying the model itself.
   - **Score calibration across groups**: Adjust predicted scores so that similar individuals from different groups receive similar calibrated probabilities.
     Example: In a recidivism risk prediction model, apply Platt scaling separately within each racial group to ensure that a 70% predicted risk corresponds to the same actual recidivism rate across groups.
   - **Threshold adjustment**: Set different decision thresholds for different groups to achieve balance in selected metrics.
     Example: In a job screening classifier, use a lower acceptance threshold for underrepresented minority candidates to achieve demographic parity in the final selection rates.
4. **Organizational and procedural measures**
   These focus on human and process elements in AI development.
   - **Diverse development teams**: Include team members with varied backgrounds, perspectives, and lived experiences.
     Example: A product team building a healthcare AI includes clinicians, ethicists, and community representatives from marginalized groups to identify potential harms early in the design process.
   - **Stakeholder consultation**: Engage affected communities and domain experts throughout development.
     Example: Before deploying a predictive policing tool, conduct workshops with community organizations in targeted neighborhoods to understand concerns and incorporate feedback into system design.
   - **Documentation of design choices**: Maintain detailed records of decisions, trade-offs, and rationales.
     Example: Produce a model card documenting the intended use case, performance across subgroups, and known limitations of a facial recognition system.
   - **Continuous monitoring**: Establish systems to track performance and fairness after deployment.
     Example: Implement automated dashboards that monitor demographic parity and error rates across groups in a real-time lending system, with alerts triggered when thresholds are violated.
5. **Technical tools**
   These are software libraries and frameworks that support bias measurement and mitigation.
   - **Fairness metrics (demographic parity, equal opportunity, equalized odds)**: Quantitative measures to evaluate model fairness.
     Example: Use demographic parity to check whether the proportion of positive predictions is similar across gender groups in a resume screening model.
   - **Bias measurement libraries (AIF360, Fairlearn)**: Open-source toolkits that implement fairness metrics and mitigation algorithms.
     Example: Use IBM's AIF360 to compute disparate impact on a credit dataset, apply the optimized

preprocessing algorithm to reweight samples, then retrain and compare fairness metrics before and after mitigation.

Sociotechnical Systems

AI systems are sociotechnical: they consist of technology interacting with people, organizations, norms, laws, and physical infrastructure.

Key components:

- **Technical layer**: algorithms, data, hardware
- **Human layer**: developers, users, affected individuals
- **Institutional layer**: companies, regulators, courts
- **Normative layer**: cultural values, ethical standards
- **Infrastructure layer**: data centers, networks, energy systems

Changes in one layer affect others. For example, a technically accurate model may produce unfair outcomes when deployed in a context with unequal baseline conditions.

Simple Ethics Frameworks in Everyday Language

1. **Utilitarianism**: Choose the action that produces the greatest overall benefit.
   This framework evaluates actions based on their consequences, aiming to maximize total well-being or happiness across all affected parties.
   Example: When deciding whether to deploy an AI system for automated hiring that increases overall efficiency but disadvantages a small group of applicants, a utilitarian would compare the total benefits (faster hiring, cost savings, broader access to jobs) against the harms (unfair exclusion of certain candidates). If the net benefit is positive, the system would be justified.
   Use case: A healthcare organization must allocate limited ventilators during a pandemic. A utilitarian approach would prioritize patients with the highest probability of survival to maximize the number of lives saved.

2. **Deontology**: Follow rules and duties regardless of consequences.
   This framework emphasizes adherence to moral rules or duties, even when breaking them might produce better outcomes.
   Example: An AI developer discovers that a model trained on scraped web data includes copyrighted material without permission. A deontologist would argue that using such data violates intellectual property rights and should not be used, regardless of how accurate or useful the resulting model would be.
   Use case: In autonomous vehicle design, a deontological approach would insist on never programming the vehicle to intentionally harm an innocent pedestrian, even if doing so would save more lives overall.

3. **Virtue ethics**: Ask what a virtuous person would do.
   This framework focuses on developing and expressing good character traits such as honesty, courage, compassion, and justice rather than following rules or calculating outcomes.
   Example: When designing a facial recognition system for public surveillance, a virtuous engineer would ask whether a compassionate and just person would create a tool that disproportionately misidentifies people of color and increases surveillance of already marginalized communities.
   Use case: A data scientist working on a predictive policing tool would consider whether a wise and

temperate professional would contribute to a system that risks reinforcing existing patterns of over-policing in certain neighborhoods.

4. **Care ethics**: Prioritize relationships and contextual needs.
   This framework emphasizes empathy, relationships, and attending to the specific needs of individuals within their particular contexts rather than applying abstract principles universally.
   Example: When developing a mental health chatbot, a care ethics perspective would focus on building trust with vulnerable users, ensuring the system recognizes when a user is in crisis and connects them to human support rather than continuing automated responses.
   Use case: In designing an AI system for elderly care monitoring, care ethics would prioritize maintaining the dignity and autonomy of the individual, respecting their personal relationships, and adapting to their unique emotional and social context.

5. **Justice-based approaches**: Focus on fairness and equitable distribution.
   This framework seeks to ensure that benefits and burdens are distributed fairly, often prioritizing the most disadvantaged groups (Rawlsian justice) or correcting historical injustices.
   Example: When allocating resources for AI research funding, a justice-based approach would direct resources toward projects that benefit underserved communities or correct existing disparities rather than solely pursuing commercially profitable applications.
   Use case: In developing an AI-based university admissions system, a justice perspective would require deliberate efforts to counteract historical exclusion by giving additional consideration to applicants from underrepresented groups or first-generation students.

Governance Tools

Algorithmic Audits

Algorithmic audits evaluate AI systems for compliance with ethical, legal, and technical standards. Audits assess fairness, transparency, accountability, robustness, and potential societal harms.

- **Internal audits**: Conducted by the developing organization or its employees.
  These audits are typically performed by dedicated internal teams or compliance departments.
  Example organization: Google has an internal Responsible AI Practices team that conducts pre-deployment and ongoing internal audits of its AI systems.
- **External audits**: Performed by independent third parties with no financial or organizational ties to the developer.
  These audits provide greater objectivity and credibility.
  Example organizations:
  - AI Now Institute (New York University) has conducted external algorithmic audits and published methodologies for public sector AI systems.
  - AlgorithmWatch, a European research and advocacy organization, performs independent audits and investigations of algorithmic systems.
  - Partnership on AI has developed frameworks and conducted external assessments of AI systems across multiple organizations.
- **Third-party certification**: Formal evaluation against established standards, leading to certification or attestation of compliance.

Certified systems can display seals or marks indicating adherence to specific criteria.
Example organizations and programs:

- ForHumanity offers ISO/IEC 42001 AI Management System certification audits through accredited third-party auditors.
- BABL AI provides third-party algorithmic impact assessments and fairness certifications.
- The IEEE CertifAIEd program evaluates AI systems against IEEE standards for ethical alignment and certifies compliance.
- DEKRA (formerly Conformity Europe) provides AI auditing and certification services aligned with emerging EU AI Act requirements.

Algorithmic Impact Assessments (AIAs)

Algorithmic Impact Assessments (AIAs) provide a structured, proactive process to evaluate the potential effects of an AI system before and during deployment to:

- Identify stakeholders
- Assess potential harms
- Evaluate benefits
- Consider alternatives
- Plan mitigation measures
- Establish monitoring

Hypothetical Scenario: AI-Based Tenant Screening System

A property management company plans to deploy an AI system to automate tenant screening for rental apartments. The system uses applicant data (credit score, income, rental history, employment status) to predict "tenant risk" and recommend approval, conditional approval, or denial.

The company conducts an Algorithmic Impact Assessment by following these structured steps:

- **Identify stakeholders**
  Relevant parties include:
  - Future tenants (especially low-income and minority applicants who may face disproportionate barriers)
  - Current and prospective landlords/property managers
  - The property management company and its developers
  - Fair housing organizations
  - Local housing authorities
  - Civil rights advocacy groups
  - Legal experts specializing in housing discrimination
- **Assess potential harms**
  Potential negative impacts identified include:
  - Disparate impact on racial and ethnic minorities due to historical credit and rental data biases
  - Exclusion of applicants with non-traditional employment (gig workers, self-employed individuals)
  - Privacy violations from collecting excessive personal data
  - Reduced housing access for people with past evictions or criminal records unrelated to future tenancy behavior
  - Reinforcement of existing neighborhood segregation patterns
  - Lack of transparency leading to distrust in the housing market

- **Evaluate benefits**
  Anticipated positive outcomes include:
  - Faster screening process, reducing vacancy periods
  - More consistent application of criteria across applicants
  - Potential reduction in human bias from individual property managers
  - Lower administrative costs for property management companies
  - Data-driven insights into factors that predict successful tenancy
- **Consider alternatives**
  Alternatives evaluated include:
  - Manual review by trained human screeners with standardized checklists
  - Hybrid system: AI provides recommendations, but humans make final decisions with mandatory review for borderline cases
  - Rule-based system using only legally permissible factors (no machine learning)
  - Open-source or publicly auditable screening model
  - No automated system (continue with the current manual process)
- **Plan mitigation measures**
  Proposed mitigations include:
  - Exclude protected characteristics and strong proxies (zip code, education level) from input features
  - Apply pre-processing debiasing techniques to balance historical outcome disparities
  - Implement human review for all denials or conditional approvals
  - Provide applicants with clear explanations of adverse decisions and appeal mechanisms
  - Conduct regular fairness audits with third-party organizations
  - Limit data retention periods and provide data deletion options
  - Establish a community advisory board with representatives from affected populations
- **Establish monitoring**
  Ongoing monitoring mechanisms include:
  - Quarterly fairness and performance reports tracking approval rates across protected groups
  - Automated alerts when demographic parity or equal opportunity metrics fall below defined thresholds
  - Annual independent external audits
  - Publicly available aggregate performance statistics
  - Formal incident response process for reported harms
  - Continuous feedback collection from applicants and tenants through anonymous channels
  - Scheduled re-assessment of the system every 12 months or after significant data or regulatory changes

Other Governance Mechanisms

Documentation standards (model cards, datasheets for datasets)

Standardized documentation formats that require developers to transparently disclose key details about models and datasets, enabling better scrutiny, reproducibility, and risk assessment.

- **Model cards**

One-page summaries describing model architecture, performance metrics across subgroups, intended use cases, limitations, ethical considerations, and fairness evaluations.

Example: For the tenant screening AI, the company publishes a model card showing approval rates by

race, income bracket, and employment type, along with known limitations (e.g., lower accuracy for gig workers) and recommended human oversight thresholds.
Reference: Model Cards for Model Reporting (Mitchell et al., 2019)
- **Datasheets for datasets**
Detailed reports covering data collection methods, sources, composition, labeling procedures, potential biases, and missing data patterns.
Example: The company releases a datasheet revealing that 68% of the training rental history data comes from urban properties in three states, notes historical eviction data biases against Black tenants, and discloses that sensitive attributes like disability status were removed during preprocessing.
Additional example: In a healthcare diagnostic model, the datasheet documents that the training images were collected primarily from hospitals in high-income countries, identifies missing demographic labels for 35% of cases, and reports higher error rates for patients with darker skin tones.
Reference: Datasheets for Datasets (Gebru et al., 2021)

Public registries of high-risk AI systems
Centralized, publicly accessible databases where organizations must register AI systems that pose significant risks to individuals or society, allowing regulators and the public to track deployment and compliance.
Example: Under the EU AI Act, the tenant screening system would be classified as high-risk (due to its impact on essential services like housing) and registered in the EU AI database. The registry entry would include the provider's name, system purpose, risk classification, and a summary of the conformity assessment.
Additional example: California's proposed Automated Decision Tool (ADT) registry would require companies using AI for housing decisions to register with the Civil Rights Department, including details on the system's purpose, data sources, and mitigation measures.
Additional example: Singapore's Model AI Governance Framework encourages voluntary registration of high-impact AI systems in finance and healthcare, with public summaries of risk assessments and governance practices.

Mandatory reporting of incidents
Legal or regulatory requirements to promptly report serious adverse events, harms, or near-misses caused by AI systems to regulators, affected individuals, or the public.
Example: If the tenant screening AI system experiences a widespread failure (e.g., a bug causes systematic denial of applications for applicants with non-traditional credit histories, affecting 2,000 low-income renters), the company must report the incident to the relevant housing regulator within 72 hours, detailing the scope of harm, root cause, and corrective actions taken.
Additional example: Under the EU AI Act (Article 73), providers of high-risk AI systems must report serious incidents that cause death, serious harm, or significant disruption to essential services, such as a medical diagnostic AI system misclassifying a life-threatening condition.
Additional example: The U.S. Algorithmic Accountability Act proposals require mandatory incident reporting for automated decision systems in housing and employment, including cases where biased outputs lead to discriminatory outcomes affecting large numbers of individuals.

Common Issues

- **Tension between fairness definitions (impossible to satisfy all simultaneously in general)**
  Different notions of fairness often conflict mathematically, as shown by impossibility theorems such as those by Kleinberg et al. (2016) and Chouldechova (2017), which demonstrate that demographic parity, equalized odds, and calibration cannot all be achieved unless base rates are identical across groups.
  Example from tenant screening scenario: Achieving demographic parity (equal approval rates across racial groups) may violate equalized odds if default rates differ historically between groups, forcing a choice between group-level equality in outcomes versus equality in error rates.

- **Difficulty measuring long-term societal impacts**
  AI systems can produce cascading effects that emerge over years, complicating quantification due to confounding variables, feedback loops, and indirect consequences. Longitudinal studies require substantial resources and face challenges in establishing causality.
  Example from tenant screening scenario: The system may initially appear fair in short-term metrics, but over time contribute to increased housing segregation by systematically denying rentals in certain neighborhoods, an impact that requires multi-year tracking of demographic shifts and economic indicators to measure.

- **Lack of standardized metrics for fairness across domains**
  Fairness metrics vary by context; what constitutes fairness in hiring (e.g., equal opportunity) may differ from healthcare (e.g., equalized odds for false negatives). No universal standard exists, leading to domain-specific adaptations and inconsistencies in evaluations.
  Example from tenant screening scenario: While demographic parity might suit measuring approval rates across income groups, individual fairness (similar individuals receive similar outcomes) could be more appropriate for applicants with similar credit profiles but different employment types, requiring custom metric selection without established guidelines for housing applications.

- **Resource constraints for small organizations to implement mitigation**
  Small entities often lack the computational power, data science expertise, or financial resources for advanced debiasing techniques, audits, or diverse teams, limiting their ability to address biases effectively compared to larger organizations.
  Example from tenant screening scenario: A small property management firm may not afford third-party audits or advanced adversarial debiasing tools, relying instead on basic pre-processing methods that inadequately address complex proxies in rental data, resulting in persistent disparities.

- **Risk of fairness washing (superficial compliance without meaningful change)**
  Organizations may adopt symbolic measures like publishing ethics statements or conducting limited audits to appear responsible, while underlying issues persist, eroding public trust and delaying substantive reforms.
  Example from tenant screening scenario: The company could release a model card claiming fairness based on a single metric (e.g., demographic parity) without addressing deeper issues like historical data biases or long-term segregation effects, using it for marketing while the system continues to disadvantage marginalized applicants.

Summary

AI systems operate within society and shape social outcomes. Ethical considerations span the entire lifecycle. Bias emerges from multiple sources and disproportionately harms marginalized groups. Sociotechnical perspectives recognize that technology cannot be separated from human and institutional elements. Mitigation requires technical, organizational, and governance approaches. Simple ethical frameworks and structured governance tools help evaluate and guide AI development toward societal benefit.