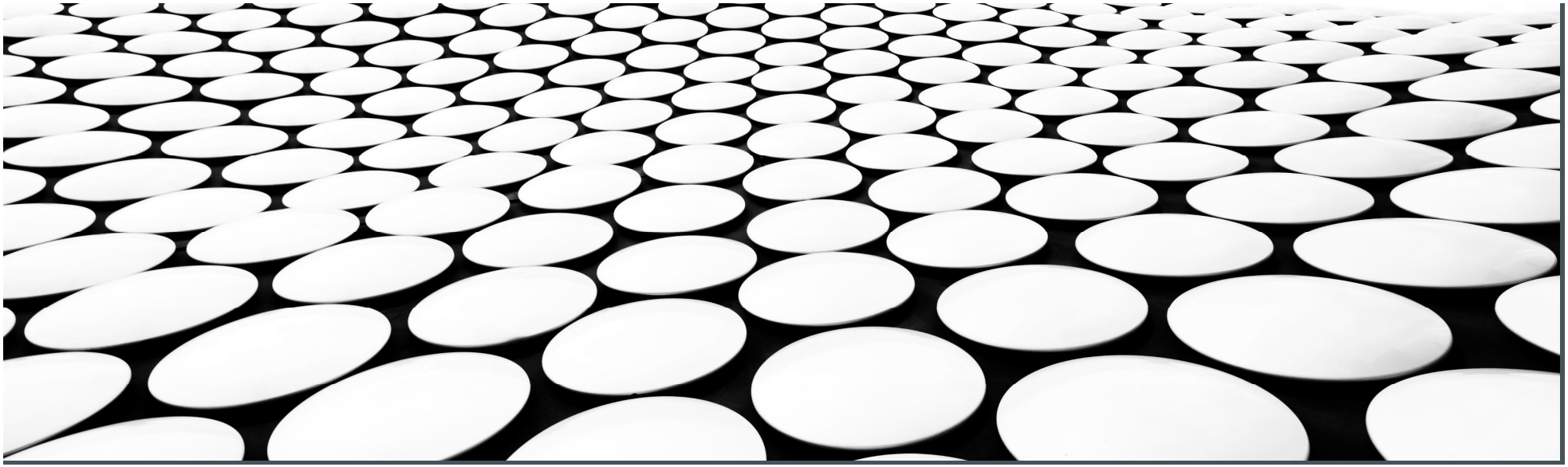
APPENDIX 2 - ADVANCED AI SECURITY – AI FAIRNESS MODELS



WHAT IS AI FAIRNESS?

- AI fairness refers to the concept of ensuring that the deployment and use of AI Systems do not result in biased outcomes or unfair treatment of individuals or groups based on certain characteristics such as race, gender, age, or other protected attributes.
  - It is important because AI systems, if not designed and implemented carefully, can perpetuate or even exacerbate existing societal biases, leading to unjust consequences in areas like hiring, lending, law enforcement, and more.

Aspect	Description
Definition of Bias	Systemic errors in AI outcomes due to prejudiced assumptions during data collection, algorithm design, or model training.
Example	AI bias in recruitment favoring candidates from overrepresented groups in training data, leading to unfair advantages and disadvantages.
Impact	Reinforces inequalities, leads to discriminatory outcomes, and affects model performance and reliability.
Challenges	Complexity in identifying and mitigating biases, given inherent biases in training data and nuanced interpretations of fairness.
Solutions	Utilization of fairness metrics such as demographic parity, equalized odds, individual fairness, counterfactual fairness, and causal reasoning.
Importance	Critical for ethical AI development, ensuring fairness, and building trust in AI systems across diverse societal applications.

BIAS IN ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Fairness Metric	Description
Demographic Parity	Requires that outcomes of an AI model are independent of sensitive attributes like race or gender, ensuring equal selection rates.
Equalized Odds	Demands equal accuracy across all groups based on sensitive attributes, ensuring consistent true and false positive rates.
Individual Fairness	Proposes treating similar individuals, similarly, based on relevant task-related similarities (e.g., credit history for loan approvals).
Counterfactual Fairness	Ensures that a model's prediction remains unchanged if a sensitive attribute is altered, maintaining fairness in alternate scenarios.
Causal Reasoning	Uses causal models to understand relationships between variables, enabling fair interventions to mitigate hidden biases and confounding factors.
Application Challenges	Addressing imbalances in data, operationalizing fairness metrics, and balancing fairness objectives with model accuracy.
Regulatory Influence	Influences regulatory frameworks like GDPR in the EU and the FCRA/ECOA in the US, emphasizing fairness in AI applications and prohibiting discriminatory practices.

FAIRNESS METRICS IN AI

AI FAIRNESS KEY CONCEPTS

Discipline	Key Concepts of Fairness
Law	Fairness includes protecting individuals and groups from discrimination or mistreatment, prohibiting biases, and ensuring decisions are not based on protected factors or social group categories.
Social Science	Fairness is often considered within social relationships, power dynamics, institutions, and markets. It addresses advantages experienced by certain groups or identities.
Quantitative Fields	Fairness is viewed as a mathematical problem, aiming to meet specific criteria such as equal allocation, equitable representation, or balanced error rates in tasks or problems.
Philosophy	Fairness is grounded in moral principles, aligning what is fair with what is morally right. Political philosophy connects fairness to broader concepts of justice and equity.

Key Concept	AI Fairness	AI Governance/Ethics
Definition	Ensures AI systems do not exhibit biases or discriminate against individuals or groups.	Encompasses principles, policies, and practices governing the development and deployment of AI.
Focus	Focuses on mitigating bias and ensuring equitable outcomes in AI decision-making.	Addresses broader ethical considerations, including transparency, accountability, and societal impacts of AI.
Goal	To promote fairness and mitigate discrimination in AI algorithms and systems.	To establish frameworks and guidelines for responsible AI development and deployment.
Methods	Utilizes metrics, algorithms, and tools to measure, detect, and mitigate biases in AI models.	Involves creating policies, codes of conduct, and regulatory frameworks for ethical AI use.
Tools	Examples include IBM AI Fairness 360 Toolkit, Google What-If Tool, qualitative fairness checklists.	Include AI ethics boards, guidelines like GDPR, frameworks like IEEE's Ethically Aligned Design.
Stakeholders	Data scientists, engineers, policymakers, and ethicists focusing on fairness in AI systems.	Governments, regulatory bodies, industry leaders, AI researchers, and civil society addressing ethical implications of AI.
Challenges	Addressing hidden biases in data, defining fairness metrics, and balancing fairness with accuracy.	Ensuring compliance with diverse global regulations, establishing universal ethical guidelines, and managing AI's societal impacts.
Impact	Mitigates discriminatory practices, enhances trust in AI, and promotes equitable societal outcomes.	Ensures AI benefits society while respecting human rights, privacy, and ethical standards.

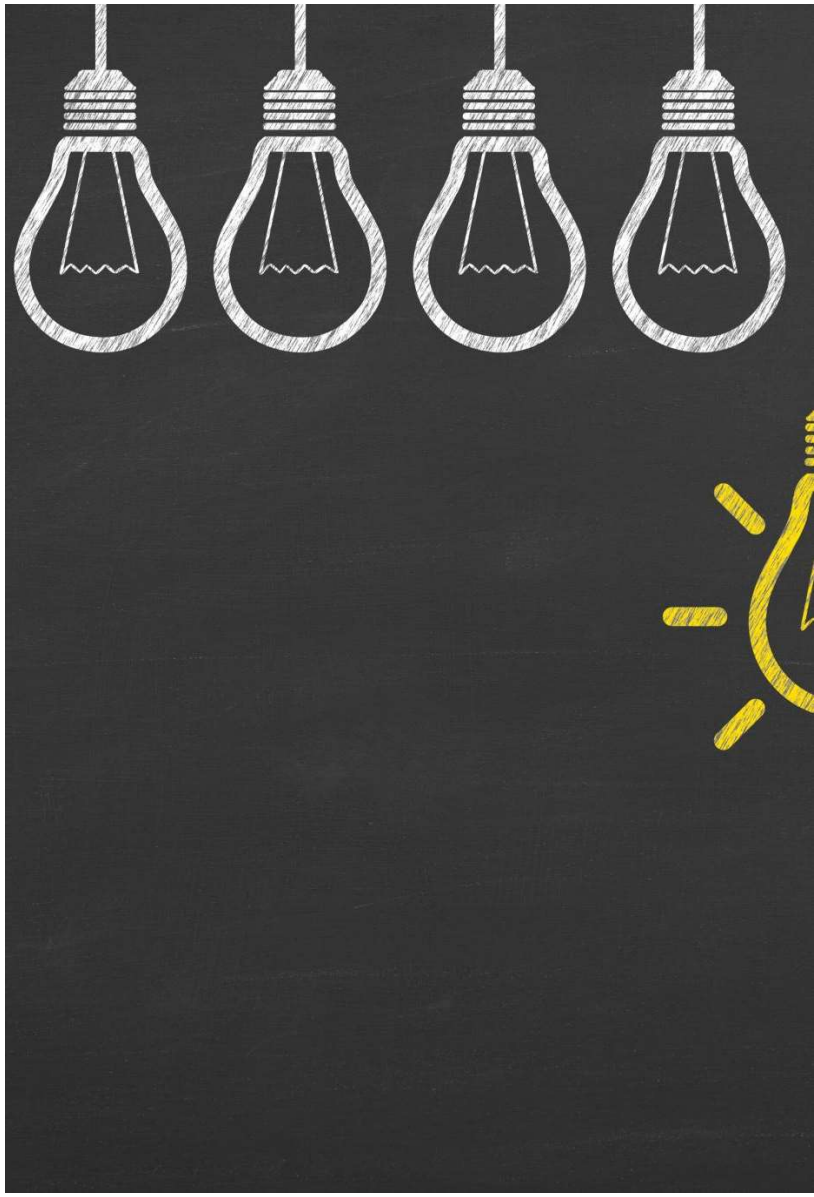
AI FAIRNESS VS. AI GOVERNANCE AND ETHICS

AI FAIRNESS

- AI fairness focuses on ensuring that AI systems do not exhibit biases or discriminate against individuals or groups based on sensitive attributes.
  - **Focus:** It concentrates on detecting and mitigating bias in AI algorithms to promote fairness and equitable outcomes in decision-making processes.
  - **Goal:** The primary goal is to create AI systems that treat all individuals fairly and reduce societal biases in automated decision-making.
  - **Methods and Tools:** Uses technical tools like fairness metrics and algorithms (e.g., AI Fairness 360) and qualitative approaches (e.g., fairness checklists) to address fairness concerns.
  - **Stakeholders:** Involves data scientists, engineers, policymakers, and ethicists collaborating to ensure fairness in AI development and deployment.
  - **Challenges:** Challenges include defining and operationalizing fairness, addressing biases in training data, and balancing fairness objectives with other performance metrics.

AI GOVERNANCE/ETHICS

- AI governance encompasses the principles, policies, and frameworks governing the development, deployment, and use of AI technologies.
- **Focus:** It addresses broader ethical considerations beyond fairness, including transparency, accountability, privacy, and societal impacts of AI.
- **Goal:** To establish regulatory frameworks, guidelines, and ethical standards that ensure AI technologies are developed and used responsibly.
- **Methods and Tools:** Includes AI ethics boards, regulatory guidelines (e.g., GDPR), industry standards (e.g., IEEE Ethically Aligned Design), and ethical impact assessments.
- **Stakeholders:** Involves governments, regulatory bodies, industry leaders, AI researchers, and civil society organizations shaping AI policies and practices.
- **Challenges:** Challenges include harmonizing global regulations, managing AI's ethical implications across diverse applications, and ensuring AI aligns with societal values and norms.



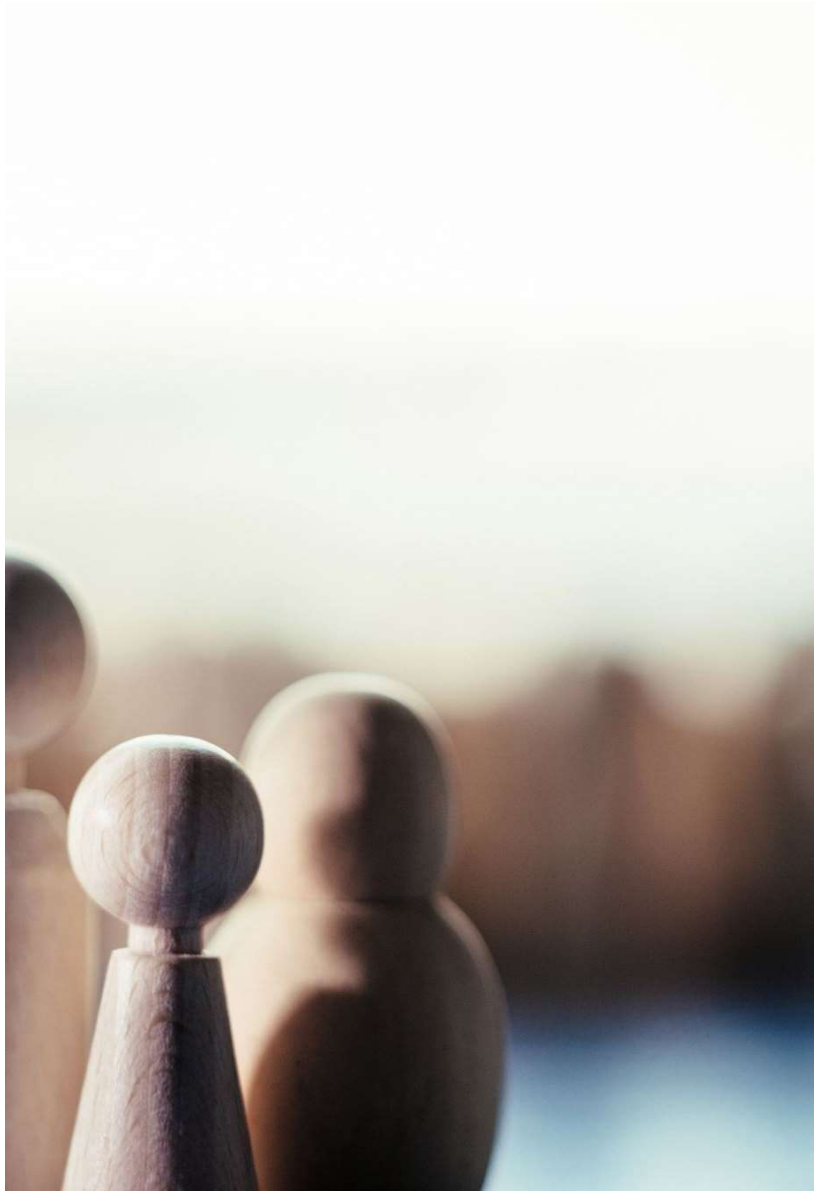
AI FAIRNESS KEY DISCIPLINES

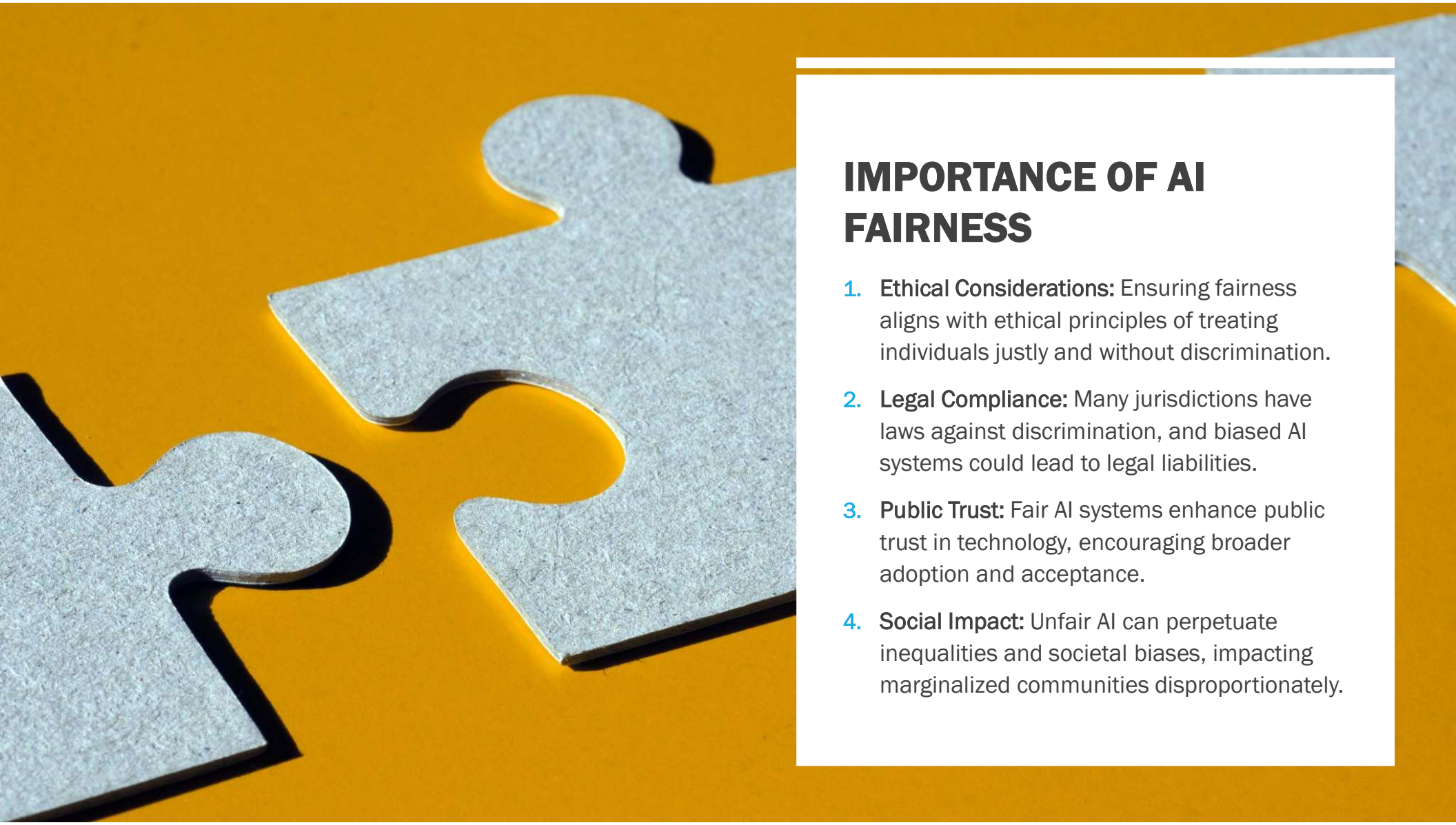
- Law: In legal contexts, fairness revolves around protecting individuals and groups from discrimination or mistreatment. Laws often prohibit biases based on factors like race, gender, or religion, ensuring decisions are impartial and based on merit rather than social identity.

- Social Science: Fairness in social sciences considers how power dynamics, institutions, and market forces influence fairness. It examines advantages or disadvantages experienced by certain groups due to societal structures and norms.

- Quantitative Fields (Math, CS, Statistics, Economics): Fairness is approached as a mathematical problem where criteria such as equal allocation of resources, equitable representation in datasets, or balanced error rates in predictive models are sought. Quantitative methods are used to measure and enforce fairness in algorithms and decision-making processes.

- Philosophy: Philosophical perspectives on fairness emphasize moral principles and the alignment of fairness with what is considered morally right. Political philosophy extends this to notions of justice and equity, exploring how fairness can be achieved in societal governance and policies.



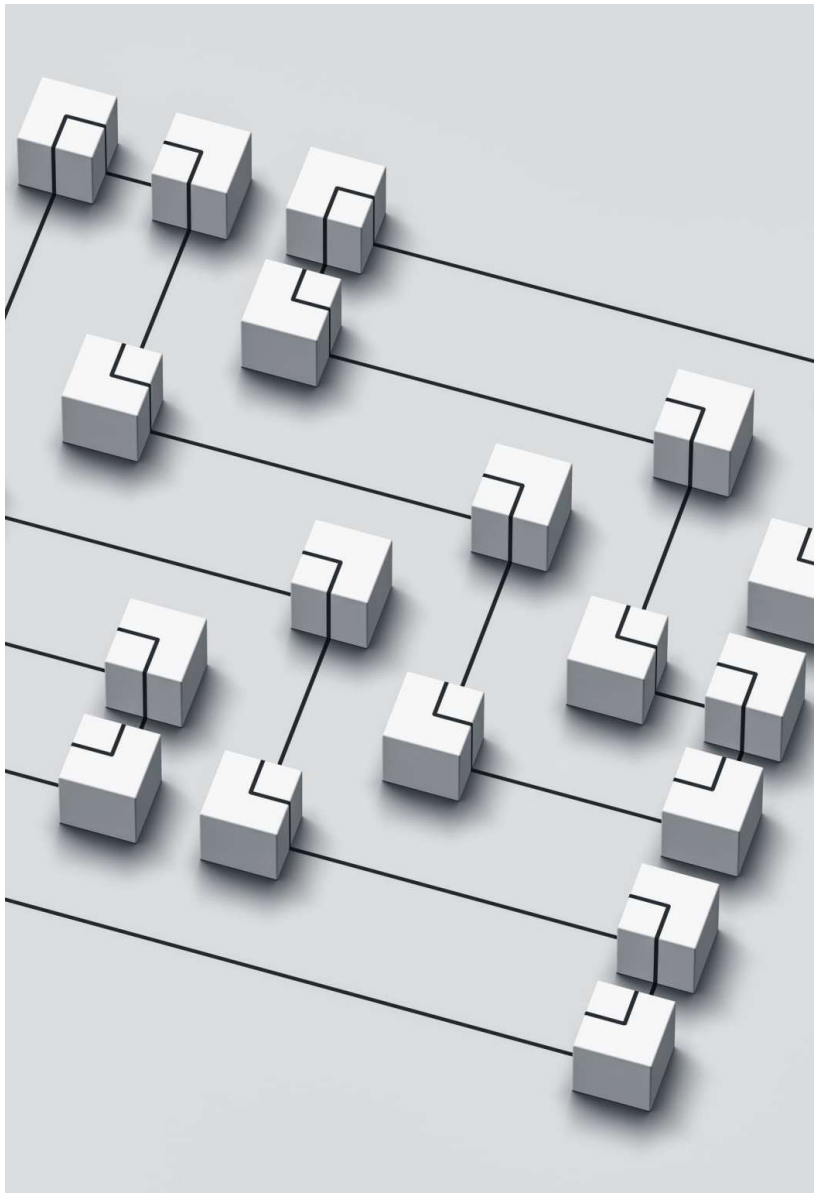


IMPORTANCE OF AI FAIRNESS

1. **Ethical Considerations:** Ensuring fairness aligns with ethical principles of treating individuals justly and without discrimination.
2. **Legal Compliance:** Many jurisdictions have laws against discrimination, and biased AI systems could lead to legal liabilities.
3. **Public Trust:** Fair AI systems enhance public trust in technology, encouraging broader adoption and acceptance.
4. **Social Impact:** Unfair AI can perpetuate inequalities and societal biases, impacting marginalized communities disproportionately.

HOW AI FAIRNESS IS ACHIEVED

1. **Data Collection and Preprocessing:** Ensuring datasets used for training AI models are diverse, representative, and free from bias.
2. **Algorithmic Design:** Developing algorithms that are inherently fair and do not discriminate based on sensitive attributes.
3. **Evaluation and Testing:** Regularly testing AI systems for fairness using metrics and techniques specific to the context and application.
4. **Post-Deployment Monitoring:** Continuously monitoring deployed AI systems to detect and mitigate biases that may emerge in real-world use.



MODELS FOR AI FAIRNESS

Models for AI Fairness: Several approaches and frameworks have been proposed to achieve AI fairness. Some prominent ones include:

1. **Fairness-Aware Machine Learning:** Techniques that modify traditional machine learning algorithms to incorporate fairness constraints during model training. Examples include adding fairness constraints to optimization objectives or adjusting model outputs to ensure fairness.
2. **Bias Detection and Mitigation:** Tools and methods for detecting biases in data and algorithms, and strategies for mitigating these biases through data preprocessing, algorithmic adjustments, or post-processing techniques.
3. **Fairness Metrics and Evaluation:** Quantitative measures to assess fairness in AI systems, such as disparate impact, equalized odds, and demographic parity. These metrics help evaluate whether AI systems are treating different groups fairly.
4. **Interpretable AI:** Models that not only make predictions but also provide explanations for their decisions can help uncover biases and ensure decisions are understandable and justifiable.
5. **Human-in-the-Loop Approaches:** Involving human oversight and intervention in AI decision-making processes to correct biases and ensure fairness in critical decisions.

KNOWN ALGORITHMS AND APPROACHES USED TO ACHIEVE AI FAIRNESS

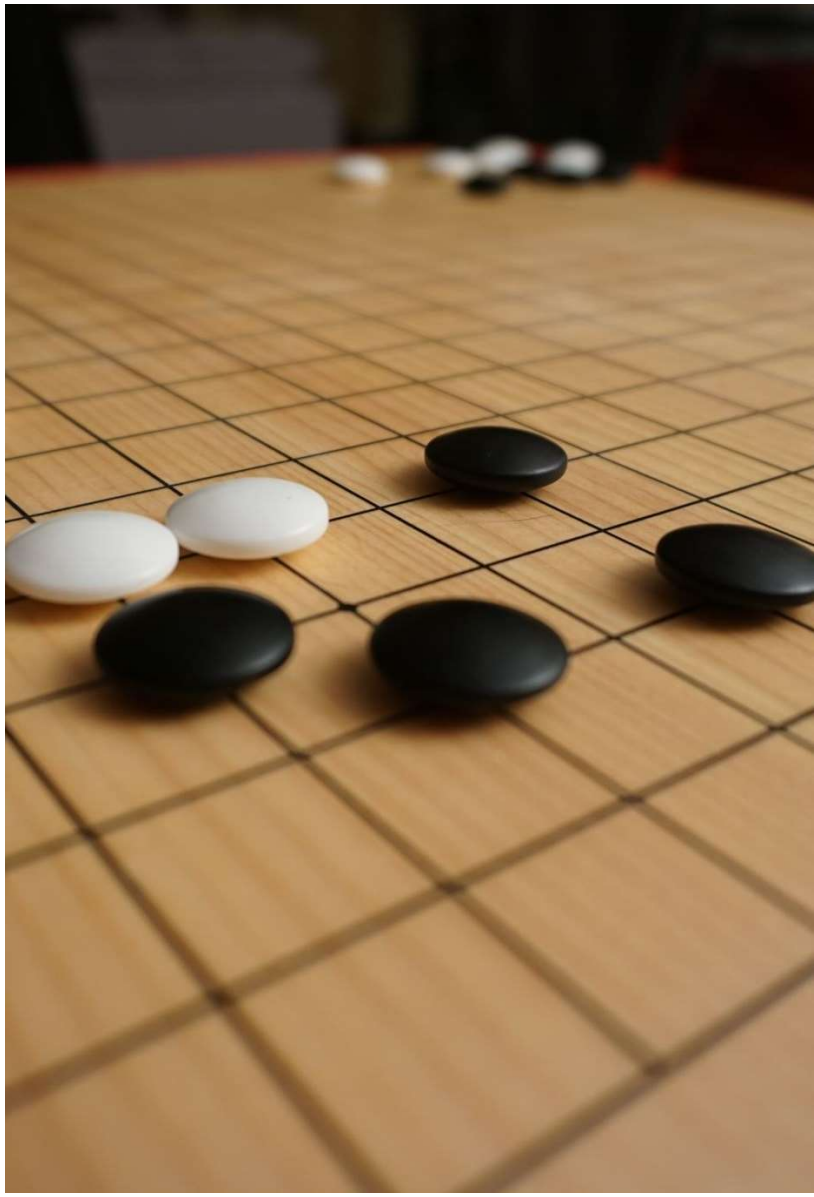
1. Preprocessing Techniques:
   - **Reweighting:** Adjusting the weights of samples in the dataset to mitigate bias.
   - **Resampling:** Techniques like oversampling or undersampling to balance dataset distributions across sensitive attributes.
   - **Subgroup Analysis:** Analyzing performance metrics across different subgroups to identify disparities.
2. In-processing Techniques:
   - **Adversarial Debiasing:** Modifying the training process by adding a debiasing objective to the learning process, often using adversarial networks to mitigate bias.
   - **Equalized Odds Postprocessing:** Adjusting model outputs to achieve equalized odds across different groups.
3. Post-processing Techniques:
   - **Calibration:** Adjusting model outputs to better align with desired fairness metrics.
   - **Threshold Adjustments:** Setting different decision thresholds for different groups to balance fairness and accuracy.
4. Regularization Techniques:
   1. **Fairness Regularization:** Adding fairness constraints to the model training process to penalize unfair decisions.
5. Algorithmic Fairness Frameworks:
   - **Fairness-aware machine learning:** Integrating fairness constraints into the optimization process of traditional machine learning algorithms.
   - **Counterfactual Fairness:** Evaluating fairness based on the idea of how outcomes would change if an individual's sensitive attribute were different.
6. Metrics and Evaluation Tools:
   - **Disparate Impact:** Measures whether there are disparities in outcomes across different groups.
   - **Equalized Odds:** Ensures that predictions are equally accurate across different groups.
   - **Demographic Parity:** Ensures that the proportion of positive outcomes is equal across different groups.
7. Interpretable AI:
   - Techniques that ensure AI models are interpretable and can explain their decisions, which helps in identifying biases and ensuring fairness.
8. Human-in-the-Loop Approaches:
   - Involving humans in the decision-making loop to review and correct biases identified by AI systems.

Aspect	AI Fairness	COMPAS Algorithm Controversy	Moving Towards Promoting Justice
Focus	Ensures AI systems do not exhibit biases or discriminate against individuals or groups.	Raised concerns about bias and fairness in algorithmic decision-making, particularly in criminal justice.	Aims to address systemic biases and inequities in societal systems, including criminal justice.
Context	Applies broadly to various AI applications where fairness is critical (e.g., hiring, lending).	Specifically focused on the criminal justice system and risk assessment in sentencing.	Seeks to reform policies and practices to achieve equitable outcomes and social justice.
Goal	Mitigates bias and promotes equitable outcomes in AI decision-making processes.	Highlighted issues of racial disparities and potential bias in automated risk assessments.	Strives for fairness, transparency, and accountability in legal and social systems.
Methods	Utilizes fairness metrics, algorithms, and tools to detect and mitigate biases in AI models.	Criticized for lack of transparency, fairness metrics, and accountability in algorithm design and implementation.	Involves legal reforms, policy changes, and community engagement to address systemic issues.
Impact	Aims to enhance trust in AI systems, reduce discrimination, and promote fair treatment.	Sparked debates and legal challenges, influencing discussions on algorithmic accountability and bias.	Seeks to rectify historical injustices, reduce disparities, and ensure equal protection under the law.
Challenges and Considerations	Challenges include defining and operationalizing fairness metrics, addressing biases in data, and balancing fairness with accuracy.	Challenges include legal and ethical dilemmas surrounding algorithmic fairness and transparency.	Challenges include overcoming resistance to change, implementing effective reforms, and measuring progress towards justice

AI FAIRNESS: COMPAS ALGORITHM CONTROVERSY, AND EFFORTS TOWARDS PROMOTING JUSTICE

SUMMARY OF FAIRNESS APPROACHES

Traditional Fairness Approach in Machine Learning & its Issues	Traditional approaches in machine learning often focus on minimizing error rates without explicitly addressing biases. Issues arise when algorithms unintentionally perpetuate or amplify existing societal biases, leading to unfair outcomes in areas like criminal justice (e.g., COMPAS algorithm controversy).
AI Fairness Tools: Technical/Quantitative	Tools like IBM's AI Fairness 360 Toolkit, Google's What-If Tool, Microsoft's fairlearn.py, and Meta's Fairness Flow focus on technical solutions. They offer metrics, algorithms, and visualizations to detect, report, and mitigate discrimination and bias in ML models, promoting fairness through quantitative analysis and adjustments.
AI Fairness Tools: Qualitative	Qualitative tools, such as co-designed AI fairness checklists and Fairness Analytics, complement technical tools by facilitating nuanced discussions about fairness. They enable teams to envision AI systems' societal roles, anticipate biases, and plan mitigations through collaborative and reflective processes.
Considerations for AI Fairness	Key considerations include early identification of fairness issues, involving diverse expertise, focusing on harm mitigation rather than perfect fairness, documenting processes and trade-offs, and using both quantitative and qualitative approaches to address biases. It's crucial to ensure transparency and accountability in AI decision-making post-development.



EXAMPLE OF AI FAIRNESS IN A LOAN APPLICATION DECISION: FROM APPROVAL TO DENIAL

- AI fairness in loan applications is a critical area where ensuring fairness is paramount to prevent discrimination and promote equitable access to financial services. Here's a summarized overview:
- AI systems are increasingly used in financial institutions to automate loan application processing, aiming to improve efficiency and accuracy. However, these systems must be designed and monitored carefully to ensure they do not perpetuate biases or unfairly disadvantage certain groups.
- Challenges:
  - **Bias in Data:** Historical data used to train AI models may reflect past discriminatory practices, leading to biased outcomes.
  - **Fairness Metrics:** Applying fairness metrics (e.g., demographic parity, equalized odds) to ensure fair outcomes without compromising predictive accuracy.
- **Example of Regulatory Compliance:** Adhering to regulations (e.g., FCRA, ECOA) that prohibit discrimination based on protected attributes such as race, gender, or ethnicity.
  - **Fair Credit Reporting Act (FCRA)** and **Equal Credit Opportunity Act (ECOA)** regulate the use of AI in credit scoring, prohibiting discrimination based on race, color, religion, national origin, sex, marital status, age, or because a person receives public assistance.
  - **The U.S. Federal Trade Commission (FTC)** has also warned companies against selling racially biased algorithms or using them in ways that might violate the law.

Key Considerations

- **Fairness Metrics in Action:** Implementing metrics to ensure that loan approval decisions are not unfairly influenced by sensitive attributes.
- **Algorithm Transparency:** Ensuring transparency in how AI systems make decisions to facilitate accountability and trust.
- **Ethical Implications:** Addressing ethical dilemmas such as trade-offs between fairness, accuracy, and profitability in loan approvals.

STRATEGIES FOR FAIRNESS

- **Data Collection:** Ensuring diverse and representative data sources to minimize bias.
- **Model Development:** Regularly auditing AI models for biases and fairness using comprehensive metrics.
- **Stakeholder Engagement:** Involving diverse stakeholders (e.g., ethicists, regulators, community advocates) in the design and deployment of AI systems.
- **Regulatory Frameworks:**
- **United States:** Laws like FCRA and ECOA regulate credit reporting and prohibit discriminatory practices in lending decisions.
- **European Union:** GDPR mandates transparency and accountability in automated decision-making processes, including loan approvals.

QUESTION AND ANSWER