Undergraduate Certificate in AI for Public Policy and Governance

AI Algorithms and Public Policy

Algorithmic Bias – Systematic and repeatable errors that favor certain groups over others in data-driven processes. Related terms: Bias mitigation, fairness, discrimination. Example: A hiring AI trained on historical employee data may downgrade applicants from underrepresented backgrounds because past hiring patterns reflected bias. Practical application: Public agencies use bias audits to assess predictive policing tools before deployment. Challenges: Identifying hidden biases requires interdisciplinary expertise; mitigation techniques (re-weighting, adversarial debiasing) can trade off accuracy for fairness, and regulatory frameworks often lag behind technical developments.

Algorithmic Accountability – The principle that creators and operators of algorithmic systems must be answerable for outcomes, including unintended harms. Related terms: Transparency, auditability, liability. Example: A city's automated traffic-signal optimization algorithm causes increased congestion in a low-income neighborhood; officials must explain the decision logic and remediate the issue. Practical application: Mandating algorithmic impact assessments (AI-IA) for high-risk public-sector deployments. Challenges: Tracing responsibility across multiple stakeholders (developers, vendors, policymakers) and establishing enforceable standards for documentation and reporting.

Algorithmic Fairness – The pursuit of equitable treatment across demographic groups when algorithms influence resource allocation, opportunity, or risk. Related terms: Group fairness, individual fairness, equity. Example: A welfare eligibility scoring system that uses income, employment history, and credit score must ensure that false-negative rates are comparable across racial groups. Practical application: Implementing fairness constraints (e.G., Demographic parity) within machine-learning pipelines used by social services. Challenges: Different fairness definitions can conflict; selecting an appropriate metric often involves value judgments and stakeholder negotiation.

Artificial General Intelligence (AGI) – A hypothetical AI capable of performing any intellectual task that a human can, with adaptability across domains. Related terms: Narrow AI, superintelligence, strong AI. Example: While current public-policy AI tools are narrow (e.G., Fraud detection), policy debates sometimes reference AGI risks such as autonomous decision-making without human oversight. Practical application: Scenario planning exercises for governments to anticipate long-term governance needs. Challenges: Uncertainty about timelines, ethical implications, and the difficulty of crafting regulation for a technology that does not yet exist.

Artificial Intelligence (AI) – The field of computer science that creates systems capable of performing tasks that normally require human cognition, such as perception, reasoning, and learning. Related terms: Machine learning, deep learning, automation. Example: Natural-language processing (NLP) chatbots used by municipal services to answer citizen inquiries. Practical application: Deploying AI-driven predictive analytics to anticipate infrastructure maintenance needs. Challenges: Balancing efficiency gains with privacy concerns, ensuring algorithmic transparency, and managing workforce transitions.

Automated Decision-Making (ADM) – The use of computer systems to make decisions without direct human intervention, often based on statistical models or rule-based logic. Related terms: Decision support systems, autonomous systems, algorithmic governance. Example: An AI system that automatically determines eligibility for housing vouchers based on income and household composition. Practical application: Streamlining benefits processing to reduce turnaround time. Challenges: Maintaining due-process rights, providing meaningful explanations to affected individuals, and preventing systemic bias.

Bias Mitigation – Techniques employed to reduce or eliminate unfair bias in data, models, or outcomes. Related terms: De-biasing, fairness-aware learning, preprocessing. Example: Using re-sampling methods to balance gender representation in training data for a criminal-risk assessment tool. Practical application: Incorporating bias-mitigation modules into the model-training pipeline of a city's child-welfare predictive system. Challenges: Mitigation may degrade predictive performance; measuring residual bias after intervention is non-trivial.

Capability Maturity Model (CMM) – A framework for assessing an organization's process maturity, adapted for AI governance to gauge readiness for responsible AI deployment. Related terms: Maturity assessment, governance framework, AI readiness. Example: A municipal IT department evaluates its data-management, model-validation, and monitoring practices against a five-level CMM scale. Practical application: Identifying gaps and prioritizing investments in AI policy infrastructure. Challenges: Customizing the model to diverse public-sector contexts and ensuring that maturity scores translate into concrete actions.

Cause-Effect Modeling – Statistical or computational techniques that infer causal relationships rather than mere correlations, supporting policy interventions. Related terms: Causal inference, counterfactual analysis, structural equation modeling. Example: Using a difference-in-differences approach to evaluate the impact of an AI-based traffic-fine policy on accident rates. Practical application: Designing evidence-based regulations that target root causes identified by AI analyses. Challenges: Data limitations, confounding variables, and the need for domain expertise to correctly specify causal models.

Certification (AI Systems) – Formal validation that an AI product meets predefined technical, ethical, and legal standards before deployment. Related terms: Compliance, standards, third-party audit. Example: A government procurement portal requires AI vendors to obtain a "Responsible AI" certificate demonstrating adherence to bias, security, and documentation criteria. Practical application: Streamlining procurement processes while ensuring accountability. Challenges: Developing universally accepted certification criteria, avoiding "checkbox" compliance, and keeping certifications up-to-date with rapid technological change.

Clear-AI – A conceptual framework emphasizing simplicity, interpretability, and transparency in AI system design for public-policy contexts. Related terms: Explainable AI, model interpretability, user-centric design. Example: Deploying a rule-based decision tree for allocating disaster relief funds, allowing officials to trace each allocation decision to specific input criteria. Practical application: Facilitating stakeholder trust and enabling effective oversight. Challenges: Balancing simplicity with the predictive power of more complex models such as deep neural networks.

Computational Social Science – The interdisciplinary study of social phenomena using computational methods, often leveraging AI to analyze large-scale digital traces. Related terms: Digital sociology, network

analysis, big data. Example: Analyzing social-media sentiment to gauge public reaction to a new AI-driven surveillance law. Practical application: Informing policy adjustments based on real-time feedback loops. Challenges: Ensuring data representativeness, protecting privacy, and interpreting algorithmic outputs in a policy-relevant manner.

Confidentiality (Data Privacy) – The obligation to protect personal information from unauthorized access, a core principle in AI-driven public-policy data handling. Related terms: GDPR, data minimization, encryption. Example: An AI system that predicts disease outbreaks must store health records in a way that prevents re-identification of individuals. Practical application: Implementing differential privacy techniques when publishing aggregated health analytics. Challenges: Achieving a useful balance between data utility for AI models and robust privacy safeguards.

Contextual Integrity – A privacy theory that assesses whether information flows align with socially accepted norms given the specific context. Related terms: Privacy, data governance, normative frameworks. Example: Using AI to share traffic-camera footage with law-enforcement agencies may breach contextual integrity if citizens expect the footage to be used only for traffic management. Practical application: Designing policy guidelines that define permissible AI-driven data uses in each public sector context. Challenges: Varying expectations across cultures and evolving norms as AI capabilities expand.

Counterfactual Explanation – A type of model interpretability that describes how minimal changes to input features could alter an AI decision. Related terms: Explainable AI, recourse, what-if analysis. Example: Providing a citizen denied a social-housing application with a statement that "if your monthly income were $200 lower, the application would be approved." Practical application: Empowering affected individuals to understand and potentially rectify adverse outcomes. Challenges: Generating realistic counterfactuals that respect legal constraints and avoid exposing sensitive data.

Crisis-Response AI – AI tools designed to support rapid decision-making during emergencies, such as natural disasters or pandemics. Related terms: Situational awareness, rapid analytics, emergency management. Example: Machine-learning models that forecast flood extents in real time to guide evacuation routes. Practical application: Integrating AI dashboards into municipal emergency operations centers. Challenges: Data sparsity under crisis conditions, ensuring model robustness, and managing public expectations about AI reliability.

Data Governance – The set of policies, standards, and processes that ensure data quality, security, and ethical use throughout its lifecycle. Related terms: Data stewardship, data ethics, data lifecycle. Example: A city establishes a data-governance board to oversee AI projects that process citizen data. Practical application: Defining access controls, retention schedules, and audit trails for AI-driven public services. Challenges: Coordinating across multiple agencies, aligning with national privacy laws, and scaling governance mechanisms for big-data environments.

Data Minimization – The principle of collecting and retaining only the data necessary to achieve a specific purpose, reducing privacy risks. Related terms: Purpose limitation, data retention, privacy by design. Example: An AI system that predicts school-dropout risk uses anonymized attendance records but does not store students' full addresses. Practical application: Designing pipelines that discard extraneous fields before

model training. Challenges: Determining the minimal dataset that still yields reliable predictions, especially when complex models benefit from richer features.

Data Provenance – Documentation of the origin, history, and transformations applied to a dataset, essential for auditability and trust. Related terms: Lineage, metadata, traceability. Example: Maintaining a provenance log that records each preprocessing step applied to crime-report data before feeding it into a predictive policing model. Practical application: Enabling regulators to verify that data sources are lawful and that transformations have not introduced bias. Challenges: Capturing provenance at scale, integrating provenance metadata across heterogeneous data platforms, and ensuring readability for non-technical auditors.

Data Sovereignty – The concept that data is subject to the laws and governance of the nation where it is collected or stored. Related terms: Jurisdiction, cross-border data flow, localization. Example: A national AI platform for healthcare analytics must store patient data within domestic data centers to comply with local regulations. Practical application: Designing cloud-architectures that respect data-localization mandates while enabling collaborative AI research. Challenges: Balancing compliance with the need for cross-jurisdictional data sharing to improve model performance.

Deep Learning – A subset of machine learning that uses multi-layered neural networks to automatically learn hierarchical feature representations. Related terms: Convolutional neural networks, recurrent neural networks, representation learning. Example: Convolutional networks that analyze satellite imagery to detect illegal logging in protected areas. Practical application: Automating environmental monitoring for policy enforcement. Challenges: High computational cost, opacity of model decisions, and the requirement for large labeled datasets.

Disparate Impact – A legal doctrine describing practices that affect protected groups differently, even without intentional discrimination. Related terms: Disparate treatment, equal opportunity, fairness. Example: An AI-driven loan-approval system that results in higher denial rates for a minority group, triggering a disparate-impact analysis. Practical application: Conducting statistical tests (e.G., The four-fourths rule) on AI outcomes before policy implementation. Challenges: Distinguishing between legitimate risk-based decisions and unlawful disparate impact, especially when historical data reflects structural inequities.

Explainable AI (XAI) – Techniques and methods that make the behavior of AI systems understandable to humans, facilitating trust and accountability. Related terms: Interpretability, transparency, model explanation. Example: Using SHAP values to illustrate how individual features contributed to a predictive model's decision to flag a business for tax audit. Practical application: Providing regulators with visual explanations during compliance reviews. Challenges: Generating faithful explanations for complex models, avoiding information overload, and ensuring explanations are meaningful to diverse stakeholders.

Fairness-Aware Machine Learning – The design of learning algorithms that explicitly incorporate fairness constraints during training. Related terms: Equitable modeling, bias-aware optimization, constrained learning. Example: Training a classifier for social-service eligibility with a constraint that false-negative rates across ethnic groups differ by no more than 5%. Practical application: Embedding fairness objectives directly into the loss function of public-policy AI tools. Challenges: Selecting appropriate fairness metrics, managing

trade-offs with accuracy, and communicating the rationale for chosen constraints to policymakers.

Federated Learning – A distributed machine-learning approach where models are trained across multiple devices or silos without centralizing raw data. Related terms: Privacy-preserving AI, edge computing, collaborative learning. Example: Multiple municipal departments collaboratively train a traffic-prediction model while keeping each department's raw sensor data on-premises. Practical application: Leveraging cross-agency data without violating data-sharing agreements. Challenges: Handling heterogeneous data quality, ensuring convergence, and protecting against model-poisoning attacks.

Feedback Loop (Algorithmic) – A cycle where AI outputs influence the data that later feeds back into the system, potentially reinforcing biases. Related terms: Reinforcement, self-fulfilling prophecy, dynamic bias. Example: A policing AI that predicts crime hotspots; increased patrols in those areas generate more incident reports, which the model then interprets as higher risk, perpetuating the cycle. Practical application: Designing monitoring mechanisms that detect and correct harmful feedback loops in public-policy AI deployments. Challenges: Diagnosing subtle loop effects, adjusting data collection practices, and establishing governance protocols for continual model re-evaluation.

Human-in-the-Loop (HITL) – An interaction paradigm where human judgment complements automated decision processes, often to ensure ethical oversight. Related terms: Oversight, decision support, hybrid intelligence. Example: An AI system flags welfare fraud cases, but a caseworker reviews each flag before any action is taken. Practical application: Reducing false positives while preserving accountability. Challenges: Designing intuitive interfaces, preventing automation bias, and allocating sufficient human resources for oversight tasks.

Impact Assessment (AI-IA) – A systematic evaluation of the potential social, ethical, and legal consequences of deploying an AI system, often required before high-risk applications. Related terms: Risk assessment, ethical review, regulatory impact. Example: Conducting an AI-IA for a facial-recognition system intended for public-space monitoring, assessing privacy, bias, and civil-rights implications. Practical application: Informing policymakers and the public about expected outcomes and mitigation strategies. Challenges: Developing standardized assessment frameworks, ensuring independent review, and updating assessments as systems evolve.

Interpretability – The degree to which a human can understand the internal mechanics or output of an AI model. Related terms: Explainability, transparency, model comprehension. Example: Choosing a logistic-regression model for credit-scoring because coefficients directly indicate the influence of each financial indicator. Practical application: Facilitating stakeholder confidence in AI-driven credit decisions. Challenges: Interpretable models may lack the predictive power of more complex algorithms; trade-offs must be carefully weighed.

Joint Data Governance Framework – A collaborative structure that brings together multiple agencies to manage shared data assets for AI projects. Related terms: Inter-agency coordination, data sharing agreements, governance board. Example: A regional transportation authority, health department, and climate office co-manage a dataset of mobility patterns to develop multimodal AI forecasts. Practical application: Aligning objectives, standardizing data definitions, and pooling resources. Challenges:

Reconciling differing legal obligations, harmonizing data standards, and maintaining joint accountability.

Knowledge Graph – A network-based representation of entities and their relationships, often used to enhance AI reasoning and policy analysis. Related terms: Semantic network, ontology, graph embeddings. Example: Building a knowledge graph linking public-housing units, demographic statistics, and crime reports to support equitable urban-planning decisions. Practical application: Enabling complex queries that combine multiple data domains for policy insights. Challenges: Ensuring data quality, updating the graph in real time, and protecting sensitive relational information.

Legitimate Expectation – A legal concept referring to the anticipation that a person's privacy or rights will be respected based on established practices or promises. Related terms: Privacy law, due process, trust. Example: Citizens have a legitimate expectation that their location data collected by city Wi-Fi sensors will not be used for unrelated surveillance. Practical application: Crafting AI policies that honor established expectations, thereby reducing legal risk. Challenges: Defining expectations in rapidly evolving technological contexts and addressing divergent expectations among different population groups.

Model Drift – The degradation of an AI model's performance over time due to changes in underlying data distributions or external conditions. Related terms: Concept drift, performance monitoring, re-training. Example: A predictive model for unemployment benefits eligibility becomes less accurate after a major economic shock, indicating drift. Practical application: Implementing continuous monitoring dashboards that trigger retraining when performance thresholds fall. Challenges: Detecting subtle drift, balancing retraining frequency with operational costs, and preserving audit trails of model versions.

Model Governance – The set of policies, processes, and controls that oversee the entire lifecycle of AI models, from development to retirement. Related terms: Model risk management, MLOps, compliance. Example: A municipal AI office establishes a model-registry that records version numbers, training data provenance, and approved use-cases for each predictive policing algorithm. Practical application: Ensuring consistent documentation, version control, and risk assessment across all AI deployments. Challenges: Integrating governance tools with existing IT infrastructure and fostering a culture of accountability among data scientists.

Neural Architecture Search (NAS) – An automated method for discovering optimal neural-network structures for a given task, often using reinforcement learning or evolutionary strategies. Related terms: AutoML, hyperparameter optimization, meta-learning. Example: Using NAS to design a lightweight model for on-device air-quality prediction in smart-city sensors. Practical application: Reducing the need for expert hand-tuning, accelerating model deployment in resource-constrained public-sector environments. Challenges: High computational cost, difficulty interpreting the resulting architectures, and ensuring that automatically generated designs meet regulatory constraints.

Open-Source AI – AI software whose source code, documentation, and often trained models are publicly available for inspection, modification, and redistribution. Related terms: Community governance, transparency, collaborative development. Example: A city adopts an open-source facial-recognition library, enabling independent security audits before deployment. Practical application: Leveraging community contributions to improve model robustness and reduce licensing expenses. Challenges: Verifying the

security of community contributions, managing divergent forks, and ensuring that open-source tools comply with local policy requirements.

Privacy-Preserving Machine Learning – Techniques that enable model training and inference without exposing sensitive data, such as differential privacy, homomorphic encryption, and secure multi-party computation. Related terms: Confidential computing, data anonymization, secure AI. Example: Training a health-risk predictor across hospitals using secure multi-party computation so that patient records never leave each institution. Practical application: Facilitating cross-institutional collaborations while respecting strict privacy regulations. Challenges: Performance overhead, selecting appropriate privacy budgets, and communicating the implications of privacy guarantees to non-technical stakeholders.

Public-Interest Algorithmic Review Board (PIARB) – An independent body tasked with evaluating AI systems for alignment with societal values, fairness, and legal compliance before public deployment. Related terms: Ethics board, oversight committee, regulatory review. Example: A PIARB reviews an AI-driven traffic-violation detection system, assessing its impact on marginalized communities. Practical application: Providing a formal checkpoint that can recommend modifications or halt deployment. Challenges: Securing diverse expertise, avoiding capture by industry interests, and ensuring timely reviews that keep pace with rapid AI innovation.

Recourse – The ability of individuals affected by an AI decision to understand, contest, and potentially alter the outcome. Related terms: Appeal, remediation, counterfactual. Example: A citizen denied a social-housing application receives a clear statement of the factors leading to denial and instructions on how to improve eligibility. Practical application: Embedding recourse mechanisms into public-service portals powered by AI. Challenges: Designing actionable feedback without revealing proprietary model details, and ensuring that recourse processes are not overly burdensome.

Regulatory Sandbox – A controlled environment where innovators can test AI applications under relaxed regulatory constraints while regulators monitor outcomes. Related terms: Pilot program, experimental governance, innovation hub. Example: A city allows a startup to trial an AI-based traffic-flow optimizer on a limited district, collecting data on safety and efficiency before broader rollout. Practical application: Accelerating responsible AI adoption while gathering evidence for future legislation. Challenges: Defining sandbox boundaries, managing liability, and preventing premature scaling of unproven technologies.

Responsible AI – An overarching framework that integrates ethical, legal, and societal considerations into AI design, deployment, and governance. Related terms: AI ethics, trustworthy AI, governance. Example: A municipal AI strategy adopts principles of fairness, transparency, accountability, and privacy to guide all AI projects. Practical application: Translating high-level principles into concrete policies, such as mandatory bias audits and public disclosure of model purposes. Challenges: Operationalizing abstract principles, aligning multiple agencies, and measuring compliance over time.

Risk Management (AI) – The systematic identification, assessment, and mitigation of potential harms associated with AI systems. Related terms: Threat modeling, mitigation strategies, governance. Example: Conducting a risk-matrix analysis for an AI-enabled public-health surveillance platform, scoring likelihood and impact of privacy breaches, bias, and system failures. Practical application: Prioritizing mitigation

actions and allocating resources accordingly. Challenges: Capturing emergent risks, integrating risk assessments into agile development cycles, and communicating risk findings to non-technical decision-makers.

Robustness – The capacity of an AI model to maintain reliable performance under varied, noisy, or adversarial inputs. Related terms: Stability, resilience, stress testing. Example: Evaluating a flood-prediction model's ability to handle missing sensor readings during extreme weather events. Practical application: Conducting robustness testing before deploying AI in critical infrastructure monitoring. Challenges: Designing comprehensive test suites, balancing robustness with model complexity, and anticipating novel failure modes.

Safeguard (AI Policy) – Specific procedural or technical measures designed to prevent or limit undesirable outcomes of AI use. Related terms: Control mechanisms, mitigation, compliance. Example: Implementing a "kill switch" that automatically disables an autonomous traffic-control system if safety thresholds are breached. Practical application: Embedding safeguards into AI contracts and operational protocols. Challenges: Ensuring safeguards are effective without undermining system utility, and updating them as threats evolve.

Scalable Governance – Governance structures and processes that can be expanded to accommodate increasing numbers of AI projects and data assets without loss of oversight. Related terms: Governance framework, capacity building, policy scaling. Example: Developing a centralized AI registry that automatically ingests metadata from departmental model-deployment pipelines. Practical application: Maintaining visibility across a growing portfolio of AI services while minimizing bureaucratic overhead. Challenges: Standardizing metadata across diverse teams, preventing registry sprawl, and ensuring that scaling does not dilute accountability.

Secure Multi-Party Computation (SMPC) – Cryptographic protocols that enable parties to jointly compute a function over their inputs while keeping those inputs private. Related terms: Privacy-preserving AI, confidential collaboration, distributed learning. Example: Several municipalities collaboratively train a crime-prediction model without revealing raw incident data to each other. Practical application: Facilitating cross-jurisdictional AI initiatives that respect data-ownership constraints. Challenges: High computational overhead, protocol complexity, and the need for specialized expertise.

Sentiment Analysis – The use of natural-language processing techniques to detect and quantify emotional tone in text data. Related terms: Opinion mining, text classification, NLP. Example: An AI system monitors social-media posts to gauge public sentiment toward a newly introduced AI-driven surveillance ordinance. Practical application: Providing policymakers with early indicators of public acceptance or resistance. Challenges: Handling sarcasm, multilingual content, and bias in language models that could misrepresent community attitudes.

Social License to Operate (SLO) – The informal approval granted by the public and stakeholders for an organization to conduct activities, based on perceived legitimacy and trust. Related terms: Public trust, stakeholder engagement, legitimacy. Example: Deploying AI-based facial-recognition cameras in public spaces may require an SLO that demonstrates community benefits outweigh privacy concerns. Practical

application: Conducting transparent outreach, impact assessments, and ongoing dialogue to maintain SLO. Challenges: Measuring intangible trust, addressing divergent community values, and reacting swiftly to breaches of the social license.

Transparency (AI) – The degree to which the inner workings, data sources, and decision logic of an AI system are open and understandable to stakeholders. Related terms: Explainability, openness, auditability. Example: Publishing a model card that details the training data, performance metrics, and intended use-cases of an AI-driven traffic-violation detection system. Practical application: Enabling citizens, auditors, and regulators to scrutinize system behavior. Challenges: Balancing transparency with intellectual-property protection, and ensuring disclosed information is comprehensible to non-technical audiences.

Uncertainty Quantification – Techniques that estimate the confidence or probability distribution around AI predictions, aiding risk-aware decision-making. Related terms: Confidence intervals, Bayesian methods, predictive uncertainty. Example: An AI model forecasting disease spread provides a 95 % confidence band, allowing health officials to plan resource allocation under uncertainty. Practical application: Incorporating uncertainty metrics into policy dashboards to inform contingency planning. Challenges: Computing reliable uncertainty estimates for complex models, communicating probabilistic information effectively, and integrating uncertainty into regulatory thresholds.

Value Alignment – The process of ensuring that AI systems pursue objectives that are consistent with human values and societal norms. Related terms: Ethical AI, alignment problem, goal specification. Example: Designing reward functions for an autonomous resource-allocation AI that prioritize equity and sustainability, reflecting public policy goals. Practical application: Conducting stakeholder workshops to define value criteria that guide model training. Challenges: Translating abstract values into concrete, quantifiable objectives, and reconciling conflicting value priorities among different stakeholder groups.

Verification and Validation (V&V) – Systematic processes for checking that an AI system is built correctly (verification) and that it meets intended purposes (validation). Related terms: Testing, quality assurance, compliance. Example: Verifying that a neural-network implementation conforms to specified architecture, then validating that the model accurately predicts traffic congestion in real-world tests. Practical application: Embedding V&V checkpoints into the AI development lifecycle for public-sector projects. Challenges: Developing domain-specific test cases, ensuring coverage of edge conditions, and maintaining documentation for regulatory review.

Vertical AI – AI solutions tailored to a specific industry or domain, incorporating specialized data, regulations, and operational constraints. Related terms: Domain-specific AI, niche AI, sectoral application. Example: A vertical AI platform for urban planning that integrates zoning codes, land-use data, and environmental impact models. Practical application: Accelerating policy analysis by providing domain-expert tools that reduce generic-AI learning curves. Challenges: Avoiding over-fitting to narrow contexts, ensuring interoperability with broader governmental data ecosystems, and updating domain knowledge as regulations evolve.

White-Box Model – An AI model whose internal logic is fully interpretable, allowing stakeholders to trace how inputs map to outputs. Related terms: Transparent model, explainable AI, rule-based system. Example:

A decision tree used to allocate disaster relief funds where each branch corresponds to a clear policy rule. Practical application: Facilitating auditability and public confidence in high-stakes allocations. Challenges: Limited expressive power compared with black-box models, and potential oversimplification of complex phenomena.