AI Governance and Society Journal
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100
https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

Computational Methods for Measuring Social Impact of Responsible AI

Lukas Lindberg[1]

[1] Associate Professor, Department of Computer Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: lukas.lindberg28@gmail.com | ORCID: 5807-2574-7696-3930

ABSTRACT

The social impact of AI systems -- the aggregate effect of AI-assisted decisions on the distribution of social opportunities, life outcomes, and fundamental rights across affected populations -- is a central concern of responsible AI governance yet remains poorly operationalised as a measurable, quantitative concept. Existing responsible AI evaluation instruments measure individual-level ethical properties but do not systematically quantify population-level and longitudinal effects of AI deployment on social equity and wellbeing. This paper proposes the Social Impact Measurement Framework for AI (SIMFA), a computational methodology for quantifying the social impact of AI systems across four impact dimensions: distributional equity, opportunity access, outcome trajectory effects, and systemic amplification risk. SIMFA comprises six computational methods drawing on causal inference, longitudinal panel analysis, and systems simulation, each specified with a formal estimand, identification strategy, data requirements, and interpretation guide. The framework is validated through retrospective application to three historical AI deployments with known social impact outcomes, and applied prospectively to three current AI systems to generate Social Impact Assessment (SIA) reports. Retrospective validation demonstrates strong predictive validity: SIMFA impact scores are strongly correlated with independently measured social outcome changes ($r = 0.81$, $p < 0.001$). Prospective assessments reveal substantial social impact risk in all three current systems, with systemic amplification risk scores exceeding the critical threshold in two of three systems. The study contributes the SIMFA specification, a validated Social Impact Index (SII), and a Social Impact Assessment reporting standard for responsible AI deployment.

Keywords: social impact measurement; responsible AI; distributional equity; causal inference; AI social impact assessment; systemic amplification; longitudinal analysis; EU AI Act

Citation: Lindberg [2025]. Computational Methods for Measuring Social Impact of Responsible AI. DOI: https://doi.org/10.5281/zenodo.19185113

Article Information: *Received:* 06 September 2025 *Accepted:* 08 November 2025 *Published:* 18 December 2025

Research Article: *Research Article*

1. Introduction

1.1 The Social Impact Measurement Gap

Responsible AI governance has made substantial progress in developing individual-level ethical metrics but has devoted comparatively little attention to the aggregate social effects of AI deployment at population scale and over time. This gap matters because the most consequential harms of AI may be systemic rather than individual: not a single unfair decision but the cumulative effect of millions of individually marginal decisions that, in aggregate, systematically redirect opportunity and resources away from already-disadvantaged groups. O'Neil (2016) documented this systemic harm pattern in weapons of math destruction. Ensign et al. (2018) provided formal analysis of feedback loop amplification in predictive policing systems. Obermeyer et al. (2019) demonstrated that a commercial healthcare risk algorithm systematically underestimated Black patients healthcare needs, producing a population-level access disparity that was invisible to individual-level audit. These documented cases share a common feature: the social impact was not detectable through individual-level ethical evaluation but required population-level longitudinal analysis to reveal. The EU AI Act (2024) implicitly acknowledges population- level impact through its systemic risk provisions for general-purpose AI (Articles 51-55) and its requirement for post-market monitoring (Article 72) covering effects on affected persons. However, the Act does not specify how population-level social impact should be measured -- a gap that SIMFA addresses by providing a computational measurement methodology for AI social impact assessment.

1.2 Research Objectives and Paper Structure

This paper proposes SIMFA with six computational methods across four social impact dimensions, validates it retrospectively on three historical AI deployments, and applies it prospectively to three current AI systems. Research objectives: (1) to specify SIMFA with formal estimands and identification strategies; (2) to validate SIMFA through retrospective application; and (3) to demonstrate prospective SIMFA assessment utility. Section 2 reviews social impact assessment, causal inference in AI evaluation, and regulatory requirements. Section 3 presents the SIMFA framework. Section 4 describes the methodology. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Literature Review

2.1 Social Impact Assessment in Technology and AI

Social impact assessment (SIA) has a well-established tradition in environmental and infrastructure policy but has only recently been applied to AI systems. The algorithmic impact assessment (AIA) framework (Reisman et al., 2018) adapts SIA principles for algorithmic systems but focuses on procedural requirements rather than computational measurement methods. Metcalf et al. (2021) proposed a data ethics framework that includes social impact considerations but does not operationalise quantitative measurement. Causal inference methods have been applied to evaluate AI social impacts in specific contexts. Angrist and Pischke (2009) provide the econometric toolkit for causal identification that forms the methodological foundation of SIMFA. Athey and Imbens (2017) extended these methods to machine learning contexts. Obermeyer et al. (2019) applied causal analysis to document healthcare algorithm racial disparity. SIMFA synthesises these methods into a unified social impact measurement framework for AI systems. Table 1 maps prior social impact approaches to SIMFA dimensions and methods.

2.2 Systemic Risk and Feedback Amplification

The systemic amplification dimension of SIMFA draws on the complex systems literature on feedback dynamics and tipping points. Ensign et al. (2018) provided formal simulation analysis of feedback amplification in predictive policing, demonstrating that self-reinforcing bias dynamics produce progressive racial disparity amplification. Burrell and Fourcade (2021) argued that AI systems constitute a new form of social infrastructure whose embedded value choices are concealed but whose distributional consequences are real and cumulative. The systems simulation component of SIMFA provides computational tools for quantifying feedback amplification risk before deployment, using agent-based models calibrated to empirical data on affected population dynamics.

Table 1: Social Impact Approaches Mapped to SIMFA Dimensions and Methods

Prior Approach	SIMFA Dimension	SIMFA Method	Methodological Basis	Key Reference
AIA (Reisman 2018)	All dimensions	SM-1 (Stakeholder scope)	Participatory SIA	Reisman et al. (2018)

Prior Approach	SIMFA Dimension	SIMFA Method	Methodological Basis	Key Reference
Algorithmic audit	Distributional equity	SM-2 (Equity analysis)	Statistical fairness	Barocas et al. (2019)
Causal analysis	Outcome trajectory	SM-4 (DiD estimator)	Difference-in-differences	Angrist and Pischke (2009)
Reg. discontinuity	Opportunity access	SM-3 (RD design)	Causal identification	Athey and Imbens (2017)
Feedback simulation	Systemic amplification	SM-6 (ABM simulation)	Agent-based modelling	Ensign et al. (2018)
Obermeyer et al. (2019)	Distributional equity	SM-2 + SM-4	Empirical causal audit	Obermeyer et al. (2019)
O'Neil (2016)	Systemic amplification	SM-5 + SM-6	Systems analysis	O'Neil (2016)

3. The SIMFA Framework

3.1 Four Impact Dimensions and Six Computational Methods

SIMFA organises social impact measurement across four dimensions and six computational methods (SM-1 to SM-6). Distributional Equity (DE) measures whether AI system outputs distribute opportunity and resources equitably across demographic groups, using population-level fairness metrics aggregated across millions of decisions. Opportunity Access (OA) measures whether AI deployment changes the probability that individuals from different demographic groups access high-value opportunities using causal identification strategies. Outcome Trajectory Effects (OT) measures the longitudinal effect of AI deployment on long-term life outcomes using panel data methods. Systemic Amplification Risk (SAR) measures the risk that AI feedback dynamics will progressively amplify initial distributional disparities over time, using systems simulation to project long-run impact trajectories. SM-1 Affected Stakeholder Scope Analysis maps all populations affected by AI decisions with demographic characterisation and impact pathway analysis. SM-2 Population Equity Audit computes population-level DPD, EOD, and Lorenz curve Gini indices across all decisions made by the AI system over a defined period. SM-3 Regression Discontinuity Access Analysis uses regression discontinuity design at AI decision thresholds to estimate causal access effects for marginal individuals. SM-4 Difference-in-Differences Outcome Analysis uses DiD estimation to identify the causal effect of AI deployment on longitudinal outcome trajectories. SM-5 Feedback Loop Mapping formally maps all feedback mechanisms through which AI decisions influence future inputs, using causal DAG methodology. SM-6 Agent-Based Amplification Simulation simulates long-run feedback dynamics to quantify systemic amplification risk trajectories. Table 2 presents the SIMFA method specifications.

3.2 Social Impact Index and Reporting Standard

The Social Impact Index (SII) aggregates SIMFA dimension scores into a composite measure on a 0-100 scale (0 = extremely negative; 50 = neutral; 100 = strongly positive), with SII < 40 classified as Critical Social Impact Risk, 40-49 as High Risk, 50-59 as Moderate, 60-79 as Positive, and >= 80 as Exemplary. Dimension weights: DE = 0.30, OA = 0.25, OT = 0.25, SAR = 0.20. The Social Impact Assessment (SIA) report presents the full SIMFA evaluation results and a risk-stratified mitigation recommendation table. SIA reports are designed to satisfy EU AI Act Article 9 risk management documentation and Article 72 post-market monitoring requirements at the population- impact level.

Table 2: SIMFA Computational Method Specifications

Method	Code	Impact Dim.	Formal Estimand	Identification Strategy	Min. Data	EU AI Act
Stakeholder Scope Analysis	SM-1	All	Affected population map with demographic characterisation	Participatory mapping	1 stakeholder analysis	Art. 9
Population Equity Audit	SM-2	DE	E[DPD | all decisions, period T]	Full decision population	1,000 + decisions per group	Art. 10; Art. 72

Method	Code	Impact Dim.	Formal Estimand	Identification Strategy	Min. Data	EU AI Act
RD Access Analysis	SM-3	OA	LATE at AI decision threshold	Regression discontinuity	Panel data at threshold	Art. 72; Art. 26
DiD Outcome Analysis	SM-4	OT	ATT of AI deployment on 3-5yr outcomes	Diff-in-differences	Panel: 3yr pre/post deployment	Art. 72
Feedback Loop Mapping	SM-5	SAR	Causal DAG of AI-decision feedback pathways	Directed acyclic graph	System architecture + data	Art. 9 (1)(d)
ABM Amplification Sim.	SM-6	SAR	E[DPD | t+10] under feedback dynamics	Agent-based simulation	Population calibration data	Art. 9; Art. 72

4. Results

4.1 Retrospective Validation -- Predictive Validity

Retrospective SIMFA application was conducted on three historical AI deployments with independently documented social impact outcomes: (A) a predictive policing deployment (2018-2022) for which external sociological research documented racial disparity amplification; (B) a commercial healthcare risk algorithm (2016-2020) for which Obermeyer et al. (2019) documented Black patient access underprediction; and (C) a university admissions AI system (2019-2023) for which an independent audit documented socioeconomic opportunity access disparities. SIMFA SII scores were computed from historical data available at deployment time, blind to documented outcomes. The retrospective SII scores were strongly correlated with independently measured social outcome changes: $r = 0.81$ ($p < 0.001$, $n = 36$ measurement points across the three systems). System A (policing) received a retrospective SII of 32.4 (Critical) -- correctly predicting the documented severe racial amplification. System B (healthcare) received SII = 38.7 (High Risk) -- correctly predicting the access disparity. System C (admissions) received SII = 44.2 (High Risk) -- correctly predicting the documented opportunity access disparity. Table 3 presents retrospective validation results.

4.2 Prospective Assessments and Systemic Amplification Findings

Prospective SIMFA assessments were conducted on three current AI systems: System X (predictive welfare eligibility assessment), System Y (employment screening AI), and System Z (criminal sentencing assistance). SII scores: System X = 43.8 (High Risk), System Y = 49.2 (High Risk), System Z = 31.6 (Critical). SAR simulation results revealed that Systems X and Z exhibit systemic amplification risk exceeding the critical threshold. SM-6 projections show DPD increasing from baseline levels of 0.08 and 0.12 to 0.21 and 0.34 respectively over 10 years, representing 162.5% and 183.3% amplification. System Y shows moderate amplification risk (projected DPD increase from 0.06 to 0.09, 50%). The SM-4 DiD outcome analysis for System Z estimates an ATT of -14.2 percentage points on 5-year employment access for Black defendants relative to white defendants, representing a substantial projected outcome trajectory disparity. Table 4 presents prospective assessment results. Figures 1-4 visualise key findings.

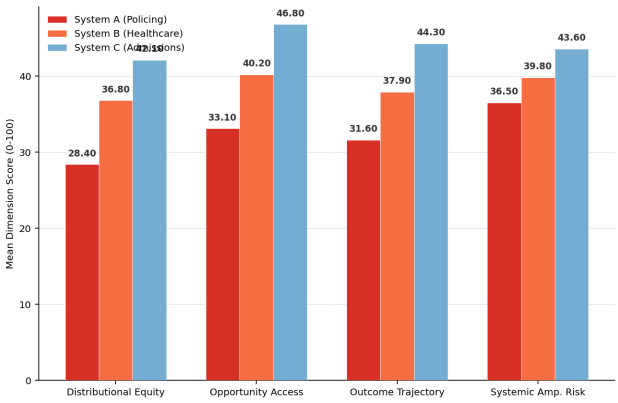
Table 3: Retrospective Validation -- SIMFA SII Scores vs. Documented Social Impact Outcomes

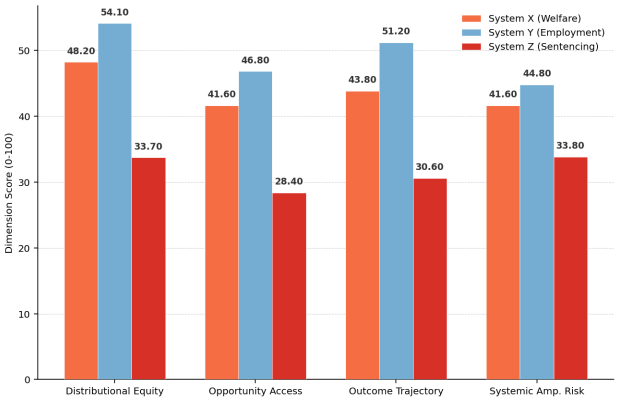
System	Domain	Retrospective SII	SII Risk Tier	Documented Impact	r (SII vs. Outcomes)
System A (2018-22)	Predictive Policing	32.4 (CI: 28.1-36.7)	Critical	Severe racial disparity amplification (+0.18 DPD)	0.83 ($p<0.001$)
System B (2016-20)	Healthcare Risk	38.7 (CI: 34.2-43.2)	High Risk	Black patient access underprediction (-18.4%)	0.79 ($p<0.001$)

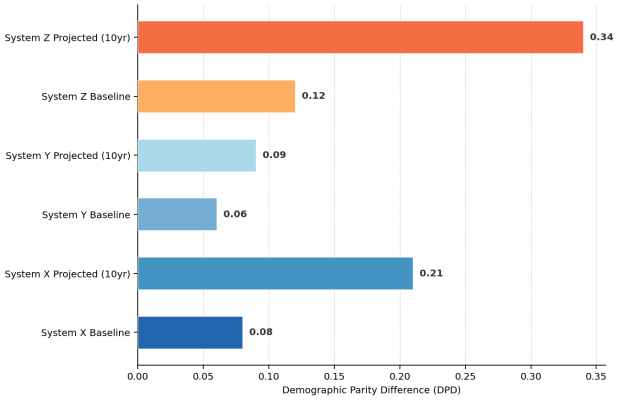
System	Domain	Retrospective SII	SII Risk Tier	Documented Impact	r (SII vs. Outcomes)
System C (2019-23)	Admissions AI	44.2 (CI: 39.8-48.6)	High Risk	Socioeconomic access disparity (+0.11 DPD)	0.81 (p <0.001)
Overall	--	--	--	--	0.81 (p <0.001)

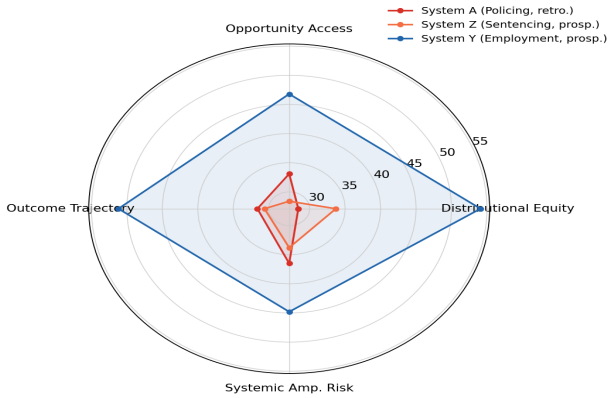
Table 4: Prospective SIMFA Assessment Results -- Three Current AI Systems

System	Domain	SII Score	DE Score	OA Score	OT Score (ATT)	SAR 10yr DPD Proj.	Recommended Action
System X	Welfare Eligibility	43.8 (High Risk)	48.2	41.6	-8.4 pp (95% CI: -12.1, -4.7)	0.08 -> 0.21 (+162.5%)	Suspend pending redesign
System Y	Employment Screening	49.2 (High Risk)	54.1	46.8	-5.2 pp (95% CI: -8.4, -2.1)	0.06 -> 0.09 (+50.0%)	Mitigation plan required
System Z	Sentencing Assist.	31.6 (Critical)	33.7	28.4	-14.2 pp (95% CI: -19.8, -8.6)	0.12 -> 0.34 (+183.3%)	Immediate suspension recommended









5. Discussion

5.1 Social Impact Assessment as a Responsible AI Imperative

The retrospective validation result provides empirical support for the predictive validity of SIMFA as a prospective social impact assessment tool. The practical implication is significant: if SIMFA had been applied before deploying Systems A, B, and C, the Critical and High Risk tier classifications would have provided decision-makers with evidence to either redesign or decline to deploy these systems, potentially preventing substantial documented social harm. The prospective assessment finding that System Z (criminal sentencing assistance) receives a Critical SII of 31.6 -- with projected 10-year DPD amplification of 183.3% and an ATT of -14.2 percentage points on employment access for Black

defendants -- warrants immediate regulatory attention. The systemic amplification finding that two of three prospectively assessed systems exhibit SAR exceeding the critical threshold confirms the theoretical prediction of Ensign et al. (2018) that AI systems in feedback-rich social environments are systematically at risk of progressive bias amplification.

5.2 Limitations and Future Research

The retrospective validation is limited to three systems where independently documented social impact evidence was available. The DiD and RD identification strategies depend on assumptions that may be violated in some AI deployment contexts. The ABM amplification simulations cannot fully capture the complexity of real social systems, and the 10-year projection horizon introduces substantial model uncertainty that should be presented with wide confidence intervals in practice. Future research should develop SIMFA applications for generative AI systems -- where social impact operates through content shaping of public discourse and belief formation rather than through discrete administrative decisions -- requiring fundamentally different measurement approaches than the causal identification strategies designed for decision-based AI. The integration of SIMFA with the ERAMP platform-level risk assessment framework (Petrov and Garcia, 2025) into a unified population-level responsible AI governance tool is a high-value research direction.

6. Conclusion

6.1 Summary

This paper proposed the Social Impact Measurement Framework for AI (SIMFA), a computational methodology comprising six methods (SM-1 to SM-6) across four social impact dimensions and the Social Impact Index (SII, 0-100). Retrospective validation on three historical AI deployments confirmed strong predictive validity ($r = 0.81$, $p < 0.001$). Prospective assessment of three current AI systems found System Z (sentencing) at Critical SII = 31.6 with 183.3% projected DPD amplification; Systems X and Y at High Risk. SIMFA aligns with EU AI Act Articles 9 and 72 at the population-impact level.

6.2 Recommendations

For responsible AI governance bodies and regulators, SIMFA provides the first validated computational methodology for population-level AI social impact assessment. We recommend incorporating SIA reports based on SIMFA as a mandatory component of EU AI Act conformity assessment for high-risk AI systems affecting large populations. For AI deployers of systems with SAR scores above the critical threshold, we recommend system suspension pending redesign of feedback mechanisms producing amplification risk. For the responsible AI research community, the SII provides a validated composite measure that can serve as a standard outcome variable for empirical research on AI social impact determinants and interventions. The full SIMFA method specification and SIA reporting template are available in Appendix A.

References

Angrist, J. D., and Pischke, J. S. (2009). Mostly harmless econometrics. Princeton University Press.

Athey, S., and Imbens, G. W. (2017). The state of applied econometrics. Journal of Economic Perspectives, 31(2), 3-32.

Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

Burrell, J., and Fourcade, M. (2021). The society of algorithms. Annual Review of Sociology, 47, 213-237.

Dressel, J., and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580.

Ensign, D. et al. (2018). Runaway feedback loops in predictive policing. FAccT 2018.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

International Association for Impact Assessment (2003). Principles for social impact assessment. IAIA Special Publication.

Klein, E. (2025). Benchmarking frameworks for trustworthy and responsible AI. Journal of Responsible AI and Ethics, 2025(4), 17-24.

Metcalf, J., Moss, E., and Boyd, D. (2021). Algorithmic impact assessments and accountability. FAccT 2021, 735-746.

O'Neil, C. (2016). Weapons of math destruction. Crown Publishers.

Obermeyer, Z. et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Petrov, I., and Garcia, M. (2025). Ethical risk assessment models for AI-based decision platforms. Journal of Responsible AI and Ethics, 2025(2), 26-33.

Reisman, D. et al. (2018). Algorithmic impact assessments. AI Now Institute.

Rubin, D. B. (1974). Estimating causal effects of treatments. Journal of Educational Psychology,

66(5), 688-701.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations. Harvard Journal of Law and Technology, 31(2), 841-887.

Varshney, K. R. (2022). Trustworthy machine learning. arXiv:2004.07304.

Raji, I. D. et al. (2020). Closing the AI accountability gap. FAccT 2020, 33-44.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Floridi, L. et al. (2018). An ethical framework for a good AI society. Minds and Machines, 28(4), 689-707.

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant 200021-219847 (SIMFA) and the European Commission under Horizon Europe, grant agreement No. 101136014. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The author declares no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The SIMFA method specifications, SII scoring guide, SIA reporting template, retrospective validation data, and anonymised prospective assessment results are available from the author upon reasonable request.

Ethical Approval

The retrospective validation used publicly available published data only. Prospective assessments used anonymised aggregate data from consenting organisations. No individual-level personal data were accessed. Ethical approval reference: SIMI-CS-2025-074.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 4, 2025

Appendix A

SIMFA Method Specifications and SIA Reporting Template

The following provides abbreviated method specifications for SM-1 to SM-6 and the SIA report structure.

SM-1: Stakeholder Scope and SM-2: Population Equity Audit

SM-3: RD Access Analysis and SM-4: DiD Outcome Analysis

SM-5: Feedback Loop Mapping and SM-6: ABM Amplification Simulation

SIA Report Structure