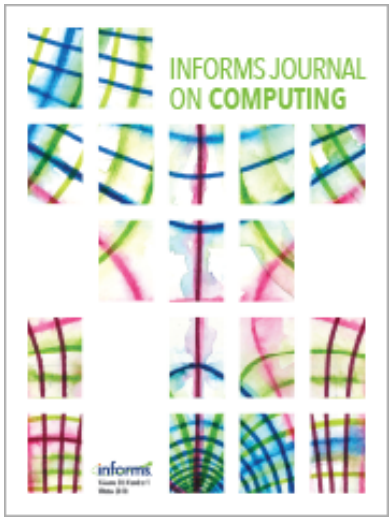


INFORMS Journal on Computing

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# Responsible AI and Data Science for Social Good

Kaushik Dutta, Ajay Kumar, Zhiling Guo, Martin Bichler, Dursun Delen, Ram Ramesh, J. Paul Brooks

To cite this article:
Kaushik Dutta, Ajay Kumar, Zhiling Guo, Martin Bichler, Dursun Delen, Ram Ramesh, J. Paul Brooks (2026) Responsible AI and Data Science for Social Good. INFORMS Journal on Computing 38(4):1033-1044. https://doi.org/10.1287/ijoc.2026.ed.v38.n4

https://pubsonline.informs.org/journal/ijoc

INFORMS JOURNAL ON COMPUTING
Vol. 38, No. 4, July–August 2026, pp. 1033–1044
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

# Responsible AI and Data Science for Social Good

**Kaushik Dutta,[a,*] Ajay Kumar,[b] Zhiling Guo,[c] Martin Bichler,[d] Dursun Delen,[e,f] Ram Ramesh,[g] J. Paul Brooks[h]**

[a] University of South Florida, Tampa, Florida 33620; [b] EMLYON Business School, Lyon, 69007 France; [c] G. Brint Ryan College of Business, University of North Texas, Denton, Texas 76201; [d] Technical University of Munich, 80333 Munich, Germany; [e] Oklahoma State University, Stillwater, Oklahoma 74078; [f] Haliç University, Istanbul, Türkiye 34060; [g] University at Buffalo, Buffalo, New York 14260; [h] Virginia Commonwealth University, Richmond, Virginia 23284

*Corresponding author

**Contact:** duttak@usf.edu, https://orcid.org/0000-0001-8076-1472 (KD); akumar@em-lyon.com (AK); zhiling.guo@unt.edu (ZG); martin.bichler@in.tum.de (MB); dursun.delen@okstate.edu (DD); rramesh@buffalo.edu (RR); jpbrooks@vcu.edu (PB)

https://doi.org/10.1287/ijoc.2026.ed.v38.n4



**Abstract.** With the rapid rise of generative artificial intelligence (AI), the responsible development and governance of AI and data science have become central concerns in both academic research and practice. As generative AI systems become increasingly embedded in daily life and play a growing role in decision making, it is essential that they operate in ethical, transparent, accountable, and socially responsible ways. In this editorial, we examine how responsible AI and data science can create meaningful societal impact across six substantive areas: judicial systems, education, communication, healthcare, bias and fairness, and interpretability. Rather than treating principles such as fairness, bias mitigation, transparency, accountability, privacy protection, robustness, interpretability, and social impact as separate organizational pillars, we view them as cross-cutting design principles that arise across these domains and methodological areas. We discuss how these principles can inform the design, deployment, and governance of AI systems that address complex societal challenges, safeguarding human values and ethical standards. Finally, we outline future research directions by contrasting pregenerative AI priorities with the emerging challenges of the postgenerative AI era. In doing so, we identify computational, methodological, and optimization frameworks that can support the responsible development and deployment of generative AI systems for meaningful societal benefit.

**Keywords:** responsible AI • generative AI • algorithmic fairness • explainability • computational and optimization models

## 1. Introduction

Artificial intelligence (AI) and data-driven decision systems are increasingly embedded in high-stakes domains such as healthcare, education, public safety, financial services, and digital communication. In these settings, algorithmic systems do not merely generate predictions; they allocate resources, shape opportunities, influence institutional processes, and affect long-term societal outcomes. As a result, questions of fairness, transparency, accountability, privacy, and social impact are no longer peripheral considerations but central design challenges.

Responsible AI is often framed primarily as an ethical imperative. Whereas normative foundations are essential, responsible AI is equally a computational and decision-theoretic problem. Designing systems that are fair, interpretable, privacy-aware, and socially beneficial requires formal models of uncertainty, constrained optimization, sequential learning, and feedback dynamics. These challenges lie at the core of operations research (OR) and computing.

From an OR perspective, fairness can be modeled through formal constraints or alternative objective functions embedded in optimization problems (Dwork et al. 2012, Hardt et al. 2016). More recent work highlights the limitations of purely static fairness constraints and emphasizes the importance of accounting for sociotechnical context and feedback effects (Selbst et al. 2019). In many applications, algorithmic decisions influence future data generation and institutional behavior, creating dynamic systems in which bias can be amplified or mitigated over time. Addressing such effects requires models of sequential decision making and dynamic optimization that incorporate fairness and societal objectives directly into learning and control policies (Lodi et al. 2024).

Similarly, transparency and interpretability are not merely communication goals but structural properties of algorithmic design. Interpretable models, subgroup-level analysis, and explanation-aware optimization require explicit trade-offs between predictive performance, robustness, and model complexity. Privacy considerations introduce additional constraints on information acquisition and data sharing, often requiring distributed or federated optimization frameworks. Across these dimensions, responsible AI becomes a problem of

balancing competing objectives under uncertainty, an area in which OR and computing provide rigorous analytical tools.

The *INFORMS Journal on Computing* is uniquely positioned to advance this agenda. The journal has long emphasized the integration of optimization, stochastic modeling, machine learning, and large-scale computational methods to address complex decision problems. Responsible AI for social good sits precisely at this intersection: it demands methodological innovation alongside careful modeling of institutional and societal environments.

Launched in May 2023, this special issue was initiated to bring together cutting-edge research examining how data-driven and AI-enabled systems can be designed, deployed, and governed to generate positive societal impact, addressing potential risks and unintended consequences. The call attracted a diverse set of submissions that spanned multiple domains of application and methodological perspectives. Across the accepted manuscripts, several themes emerged, including applications in judicial systems, education, communication, and healthcare, alongside foundational concerns central to responsible AI such as bias and fairness and interpretability. Collectively, these themes highlight both the breadth of social impact domains and the shared technical and ethical challenges that must be addressed to advance AI systems that are transparent, explainable, and socially beneficial. Although the application areas vary, the papers share common themes. First, they move beyond purely predictive modeling toward decision-aware and system-level frameworks. Second, they explicitly confront trade-offs between performance and societal objectives, often through constrained optimization, dynamic learning, or structured evaluation metrics. Third, they illustrate how algorithmic design choices shape long-term system behavior and social outcomes.

Collectively, the contributions underscore that responsible AI cannot be achieved through post hoc adjustments alone. Instead, societal considerations must be integrated into data collection processes, model design, exploration policies, deployment strategies, and governance mechanisms. We hope that this special issue encourages further research that treats fairness, transparency, and social impact not as external constraints imposed on otherwise optimal systems but as intrinsic design objectives within computational decision frameworks.

## 2. Judicial Systems

OR and data-driven decision techniques have been applied globally to improve judicial systems in two critical areas: efficiency (timely action) and fairness (equity and impartiality in outcomes). Methodological approaches ranging from traditional operations research techniques such as mixed integer programming (Shahabsafa et al. 2018), queuing theory (Bakshi et al. 2025), and stochastic modeling (Blumstein 2007) to modern digital technologies, including AI (Taruffo 1998, Freeman et al. 2026), have been applied to make society safe and apply a proper justice system.

One major social challenge in courts worldwide is excessive delay, which undermines access to justice. Optimization, queuing theory, and simulation—classic OR methods—have been applied to court case scheduling and flow management with striking results (Brooks 2012, Bakshi et al. 2025). Researchers note that "Justice works best when it is both fair and timely." OR offers a practical blueprint to achieve that timeliness, which is a social good for victims, defendants, and the public (Bakshi et al. 2025, Informs 2025). The paper "Integrating Mental Health and Juvenile Justice Outcomes: A Case Study of Model-Agnostic Interpretable Machine Learning" extends the existing research on efficiency in judicial systems, but it does more than that by focusing on a vulnerable group: children. By integrating mental health in the juvenile justice system with the help of AI, the research achieves a better outcome. The social implication of this unique research is unparalleled in nature. We hope this research encourages further work on applying OR and machine learning techniques to make the juvenile system more efficient and positive-outcome driven.

OR has been applied to tackle urban and organized crime. Freeman et al. (2022) develop models to scrape and analyze online sex advertisement data. This data-driven approach leads directly to enhanced police sting operations. The study demonstrates how OR and analytics can uncover hidden criminal networks and guide effective interdiction of human trafficking. Dimas et al. (2022) is an excellent survey of academic research on the application of OR and analytics to address human trafficking. Konrad et al. (2023) present domain-specific themes grouped around the representation of human trafficking and the consideration of survivors and communities to guide operations research and analytics practitioners in conducting responsible, nuanced antitrafficking research.

Levine et al. (2017) describe the New York Police Department's Domain Awareness System, a citywide integrated analytics platform. The system fuses feeds from sensors, databases, closed-circuit television cameras,

license plate readers, and other sources into a central hub that pushes tailored alerts to officers' smartphones and precinct dashboards. Machine learning models detect crime patterns and potential threats (including terrorism indicators), enabling proactive deployment of resources. Chohlas-Wood and Levine (2019) develop Patternizr, an OR-driven recommendation engine that automates the search for related crimes in police databases. Patternizr's models compute a similarity score between any two crime incidents based on dozens of features (MO, time, location, etc.). When an investigator provides a new "seed" case, Patternizr scans hundreds of thousands of records to recommend the most similar past incidents, essentially suggesting crimes likely committed by the same offender. Jain et al. (2010) develop decision support systems that use OR to randomize security patrols intelligently. The software computes an optimal randomized patrol schedule that maximizes security coverage weighted by target importance.

The paper "Guardians of Tomorrow: Leveraging Responsible AI for Early Detection and Response to Criminal Threats" in this special issue is built on top of this existing research. The research makes the movie *Minority Report* by Steven Spielberg, produced in 2002, a reality. In the movie, a crime is predicted and intervened before the crime has happened. This paper develops a similar approach to detecting a criminal threat so that it can be prevented for social good purposes.

OR methods may offer unrealized improvements in the criminal and juvenile justice systems. A few key problems in this area to apply OR techniques include building full cost–benefit models of sentencing that account for individual and community impacts; understanding how pretrial detention compounds disadvantage for marginalized groups; correcting crime prediction systems trained on historically biased data; improving prediction accuracy to ethically guide early intervention; and optimizing surveillance deployment, accounting for privacy costs. In juvenile justice specifically, OR tools are needed to prevent low-risk youth from being unnecessarily drawn into the system, to ensure that risk assessment tools are used consistently, and to match adjudicated youth to the right rehabilitative resources at the right time. The papers in this special issue on this topic set the stage for further research in these areas.

## 3. Education

The deployment of AI in education has evolved from early efforts focused primarily on automation and predictive modeling toward a broader sociotechnical perspective encompassing multistakeholder support, AI literacy, human–AI collaboration, and governance considerations related to pedagogy, equity, and ethics. This shift is catalyzed by the rise of learning analytics and educational data mining that demonstrate both the promise of large-scale measurement for improving learning outcomes and the ethical risks associated with intensive data capture, including concerns about privacy, surveillance, and accountability (Siemens 2013). In response, the AI-in-education community has developed ethics frameworks that emphasize learner agency, equity, and responsible design aligned with educational values and institutional contexts (Porayska-Pomsta et al. 2023). Complementing these developments, syntheses grounded in the fairness, accountability, transparency, and ethics framework underscore the need to evaluate AI systems not only for technical performance but also for their broader implications for equity, accountability, and institutional impact (Memarian and Doleck 2023). At the policy level, international guidance further reinforces this trajectory by advocating governance mechanisms that ensure AI-driven innovation advances inclusion and social good (Miao et al. 2021). Two special issue papers exemplify this emerging responsible AI orientation.

The study "CAAC: Coattentive Actionability Classification for Assessing Patient Education Videos" develops a transformer-based hierarchical coattention framework to automatically assess the actionability of patient education videos according to the Patient Education Materials Assessment Tool guidelines. Grounded in cognitive theory of multimedia learning (CTML), the architecture is explicitly designed to capture cross-modal coherence and temporal sequencing, key determinants of whether viewers can identify and execute recommended actions. By embedding pedagogical theory and domain-specific standards into a multimodal modeling framework, CAAC demonstrates how scalable AI systems can provide guideline-consistent quality assessment of educational health content.

The paper "Quantifying the Academic Quality of Children's Videos Using Machine Comprehension" proposes a scalable framework to evaluate the academic quality of children's YouTube videos by measuring their alignment with school textbook curricula. The framework integrates transcript encoding with visual frame-caption representations, drawing conceptually on CTML and formative assessment principles to justify comprehension-based evaluation. By introducing a weighted academic quality metric for ranking channels,

the study illustrates how machine comprehension can operationalize curriculum-aligned quality assessment at a platform scale.

Taken together, these two studies exemplify a shift in responsible AI in education from building student-facing adaptive tools toward developing system-level mechanisms for scalable educational quality assurance. By operationalizing pedagogical standards into computational architectures that evaluate instructional value at scale, both approaches emphasize multimodal coherence, theory-grounded design aligned with domain standards, transparency of evaluation criteria, and support for institutional oversight. In doing so, AI is repositioned as a governance-oriented infrastructure that enhances accountability, consistency, and educational quality across digital platforms. Future research should further integrate technological innovation with governance and policy considerations to ensure that responsible AI in education delivers measurable educational benefits and contributes meaningfully to social good.

## 4. Communication

The proliferation of social media has fundamentally reshaped the digital communication landscape, dramatically increasing both the speed and scale of information exchange (Gillespie 2018, Van Dijck et al. 2018). Alongside these benefits, however, harmful phenomena, such as hate speech and misinformation, have emerged as pressing societal challenges (Lazer et al. 2018, Vosoughi et al. 2018). Early research on automated content moderation largely emphasizes predictive accuracy, prioritizing improvements in classification performance through increasingly sophisticated machine learning and deep learning architectures. Yet real-world deployment of these systems has exposed critical limitations and vulnerabilities that extend beyond pure predictive performance, including adversarial vulnerability, distributional shifts, bias amplification, and overconfidence. As a result, responsible AI in communication has progressed along several major dimensions.

### 4.1. From Accuracy to Robustness

Early research on hate speech and misinformation detection focuses primarily on improving classification accuracy. However, as real-world adversaries actively attempt to evade detection, robustness has become a central concern (Madry et al. 2017). Contemporary work addresses adversarial manipulation at the character, word, and multimodal levels, recognizing that attackers exploit both linguistic ambiguity and cross-modal inconsistencies. Techniques such as adversarial training, attack modeling, and robustness-aware representation learning are increasingly integrated into communication AI systems. Robustness is now treated as a primary design objective rather than a secondary evaluation metric, reflecting a shift toward reliability under strategic manipulation (Gorwa et al. 2020).

### 4.2. From Unimodal to Multimodal Intelligence

Misinformation campaigns frequently exploit inconsistencies between text and images, visual manipulation, and contextual cues such as publisher metadata (Shu et al. 2017). Consequently, multimodal fusion approaches—integrating text, images, and auxiliary metadata—have become mainstream (Baltrušaitis et al. 2018). Current research emphasizes principled modality weighting and cross-modal alignment instead of naïve feature concatenation, thereby enhancing semantic coherence and improving detection reliability.

### 4.3. From Prediction to Trustworthy AI

When deployed in high-stakes communication environments such as elections, public health crises, or social unrest, AI systems must provide not only accurate predictions but also reliable confidence estimates (Floridi et al. 2018). Trustworthiness is now understood as a multidimensional construct encompassing robustness, calibration, interpretability, fairness, and domain adaptability (Glikson and Woolley 2020).

### 4.4. Early Detection and Temporal Adaptation

Misinformation evolves rapidly, causing static models to degrade under temporal and contextual shifts (Quiñonero-Candela et al. 2009). Recent advances focus on domain adaptation, adversarial transfer learning, temporal robustness modeling, and early stage detection under limited propagation signals (Vosoughi et al. 2018). These developments reflect a broader transition from static classification systems to adaptive, resilient communication AI frameworks capable of responding to dynamic adversarial environments.

The paper "Semantic Aggregated Adversarial Training (SAAT) for Hate Speech Detection" addresses a critical limitation in hate speech detection: ensuring adversarial robustness under conditions of class imbalance and

semantic ambiguity. The authors propose SAAT, a framework that integrates word embedding–level adversarial training, class-balanced loss reweighting, and a novel semantic aggregation regularization designed to enhance feature separability between hateful and nonhateful content. By jointly addressing adversarial vulnerability, class imbalance, and feature inseparability, SAAT enhances robustness without sacrificing classification accuracy. The study contributes to responsible AI in communication by ensuring the reliability and resilience of moderation systems in adversarial online environments.

The study "Confidence-Aware Multimodal Learning for Trustworthy Fake News Detection" develops a trustworthy multimodal fake news detection framework that addresses two overlooked issues: effective multimodal semantic utilization and confidence miscalibration. The authors propose a semantic similarity–based multimodal fusion mechanism that dynamically assigns weights based on modality contribution. They further introduce a theoretically grounded post hoc temperature-scaling calibration method that improves reliability without altering classification accuracy. The study advances trustworthy AI in communication by ensuring not only accuracy but also reliable confidence calibration.

The paper "Combating Fake News on Social Media: An Early Detection Approach Using Multimodal Adversarial Transfer Learning" focuses on early stage fake news detection under temporal shifts. The proposed MATRAL framework integrates multimodal feature extraction with adversarial transfer learning to address the challenge that early detection lacks propagation signals and that fake news tactics evolve over time. Through adversarial domain adaptation, MATRAL learns temporally transferable representations. The paper contributes to responsible AI in communication by enhancing adaptability and robustness in dynamic communication environments.

Together, these contributions reflect a broader paradigm shift from static, accuracy-maximizing classifiers to resilient, trustworthy, and adaptive AI systems for communication. Future research should extend beyond model-level performance gains to examine how responsible AI architectures reshape communication ecosystems at both platform and societal levels. Key directions include investigating behavioral trust calibration, the incentive compatibility of responsible AI deployment, advances in causal and interpretable multimodal reasoning, and longitudinal analyses of how adaptive detection systems influence misinformation diffusion and public trust over time. Addressing these issues requires interdisciplinary integration across AI, communication theory, behavioral decision science, and platform economics to ensure that technical innovations translate into measurable social good and effective governance outcomes.

## 5. Healthcare

Responsible AI is connected to the healthcare domain in several interrelated ways as this field relies heavily on data and involves significant ethical considerations in the use of AI (Koski et al. 2025). When designing responsible healthcare monitoring systems, it is essential to ground their architecture and deployment in core principles such as fairness and bias reduction, transparency and explainability, accountability, safety and reliability, data privacy, societal impact, and the incorporation of human oversight (Koski et al. 2025). Healthcare monitoring systems, such as wearable devices, intensive care unit monitoring tools, and remote patient monitoring platforms, are a rapidly growing area for AI deployment in healthcare, and they offer a direct, practical context for applying responsible AI principles within the healthcare ecosystem (Trocin et al. 2023). These systems continuously collect and analyze real-time physiological and behavioral data and help early detection, risk prediction, and timely clinical intervention, but at the same time, it is very important to ensure that such data-driven care improves patient outcomes, maintaining trust and upholding ethical standards (Bawack et al. 2025). Samorani et al. (2022) demonstrate that machine learning–driven appointment scheduling systems cause Black patients to wait approximately 30% longer than non-Black patients and propose a race-aware optimization objective that eliminates this racial disparity, maintaining scheduling efficiency comparable to state-of-the-art methods.

In this special issue, we publish two papers that address two critical challenges in the healthcare domain: developing responsible AI–based models for infodemic potential prediction and privacy-preserving smart mobile health monitoring. The paper "A Novel Personalized Federated Learning Method for Privacy-Preserving Smart Mobile Health Monitoring" develops a federated learning–based model to monitor patients' health status effectively, addressing the limitations of conventional methods that often fail to adequately protect patient privacy. In their study, the authors explore how mobile technologies and AI can be leveraged to enable privacy-preserving health monitoring and promote greater social good. They conduct extensive experiments on real-world health data sets, demonstrating that their proposed model outperforms traditional approaches in both predictive accuracy and privacy protection.

The paper "Responsible AI-Enabled Infodemic Management: A Hypergraph-Based Infodemic Topic Prediction Framework" develops a four-phase framework for infodemic potential prediction to responsibly address this challenge as infodemics during public health events can generate widespread misinformation, public panic, weakening of trust in institutions, and misguided health behaviors that could lead to severe societal and healthcare consequences. Their proposed model advances the field of graph learning by introducing a novel directed hypergraph modeling and transformation approach. The study presents a four-phase prediction framework and evaluates its performance using data from two real-world public health events: the COVID-19 pandemic and the monkeypox outbreak. The results demonstrate the advantages of their method over existing approaches. Overall, this research contributes to responsible infodemic management by offering a topic-level framework that enables the monitoring and mitigation of infodemics in a human-centric, informative, and globally informed manner.

When discussing future research directions for developing responsible AI and social good–based predictive and optimization models in the healthcare domain, the focus should be placed on integrating ethical, societal, and technical considerations throughout the model life cycle. A promising direction is the development of privacy-preserving and fairness-aware frameworks that could balance predictive accuracy with the protection of patients' sensitive data and the mitigation of algorithmic bias. Researchers could also investigate multimodal and multisource data integration in the healthcare domain, combining clinical, social, and behavioral data to enhance the accuracy, robustness, and societal relevance of predictive healthcare models. Another key area in the era of generative AI is the development of explainable and transparent predictive and optimization models for healthcare as these could enable stakeholders, including clinicians, policymakers, and patients, to understand, trust, and effectively act on generative AI–driven insights and recommendations. Additionally, the development of adaptive and context-aware models that can respond to dynamic social, environmental, and public health conditions will significantly enhance the real-world impact of AI for social good in healthcare. Finally, interdisciplinary collaborations among AI researchers, social scientists, clinicians, and other domain experts are essential for ensuring that predictive and optimization models not only perform well technically but also maximize societal benefit and ethical responsibility.

## 6. Bias and Fairness

Bias and fairness have emerged as central concerns in the deployment of AI and data-driven decision systems in socially sensitive domains. Algorithmic bias can arise from historical inequities embedded in data, selective or censored feedback, model misspecification, or strategic responses by decision subjects. When left unaddressed, such biases can lead to systematic disparities across demographic groups in high-stakes settings such as education, lending, hiring, healthcare, and online platforms. From an operations research and computing perspective, fairness concerns are tightly linked to classical questions of decision making under uncertainty, dynamic learning, resource allocation, and system-wide performance but with the additional requirement that outcomes be equitable across protected groups (Selbst et al. 2019, Barocas et al. 2023).

Early work on algorithmic fairness in OR and related fields focuses on defining and measuring fairness through formal constraints, such as demographic parity, equalized odds, or equal opportunity, and studying the efficiency–fairness trade-offs induced by these constraints (Dwork et al. 2012, Hardt et al. 2016). Subsequent research highlights the dynamic nature of bias, showing how learning algorithms interact with endogenous data generation processes and feedback loops, potentially amplifying initial disparities (Selbst et al. 2019). Recent studies have, therefore, shifted attention from purely static fairness constraints toward dynamic, system-level perspectives that account for learning, exploration, and long-term impacts on different populations (Lodi et al. 2024). This evolution naturally positions OR as a key discipline for advancing fairness-aware AI given its emphasis on optimization, dynamics, and rigorous performance guarantees.

The three papers in this special issue reflect this broader shift and illustrate complementary approaches to bias and fairness across different application contexts and methodological lenses.

The paper "Systemic Fairness & College Admissions" studies fairness not at the level of a single decision maker but at the level of an entire admissions ecosystem. Using agent-based simulation grounded in real-world data, the authors show how heterogeneity in AI sophistication, selection criteria, and application constraints across institutions can generate systemic inequities even when individual decision rules appear reasonable. By extending classical group fairness metrics to the system level and introducing the notion of choice parity, this work highlights the importance of modeling multiagent interactions and institutional diversity when evaluating fairness outcomes.

The paper "Adaptive Bounded Exploration and Intermediate Actions for Data Debiasing" addresses a different but equally fundamental source of unfairness: biased and censored training data. Focusing on sequential decision problems with costly and censored feedback, the authors propose bounded exploration algorithms that actively guide data collection to mitigate statistical bias, controlling exploration costs. By explicitly modeling the trade-off between debiasing speed, decision accuracy, and incurred costs, this work connects fairness concerns to classical OR themes such as exploration–exploitation trade-offs, stochastic control, and dynamic optimization. Importantly, the paper demonstrates how data debiasing mechanisms can complement rather than substitute for fairness constraints imposed at the algorithmic level.

The paper "Mitigating Age-Related Bias in Large Language Models" extends the fairness discussion to modern foundation models. The authors study age-related bias in large language models and propose a two-stage, posttraining bias mitigation framework that combines reinforcement learning, human-in-the-loop feedback, and agent-based debate mechanisms. By focusing on output-level bias mitigation without modifying model parameters, this work opens new directions for fairness interventions in large-scale AI systems in which retraining may be infeasible. The paper also illustrates how fairness in generative models requires different tools than traditional predictive settings yet still benefits from formal evaluation metrics and structured decision frameworks.

Together, these papers point to several promising directions for future research at the intersection of operations research, computing, and fairness. One important avenue is the development of system-level fairness models that explicitly capture interactions among multiple decision makers, institutions, and learning agents, moving beyond single-algorithm analyses. Another is the integration of fairness considerations into sequential decision-making and learning models, in which exploration policies, data acquisition, and feedback mechanisms can be designed to reduce bias over time, respecting cost and risk constraints. A third direction concerns fairness in foundation models and other large-scale AI systems, in which OR tools such as dynamic programming, incentive design, and robust optimization could inform principled postprocessing and governance mechanisms.

More broadly, these contributions underscore that fairness should not be viewed as an external constraint imposed on otherwise optimal systems but as a core design objective that shapes how data are collected, models are trained, and decisions are made over time. We hope this special issue encourages further OR-driven research that treats bias and fairness as dynamic, systemic, and optimization-relevant phenomena and that develops analytically grounded methods capable of delivering both high performance and socially responsible outcomes.

## 7. Interpretability and Explainability

The growing emphasis on responsible AI and data science for social good has brought interpretability and transparency to the forefront of computational research. As highlighted in foundational works on explainable machine learning, the tension between predictive accuracy and human interpretability has long characterized advanced analytics: increasingly powerful models often operate as opaque black boxes, limiting trust, accountability, and effective deployment in high-stakes domains, such as healthcare, public policy, and social services (Delen 2020). Rudin (2019) offers a distinction between interpretability and explainability in which interpretability is taken to be an inherent property of a model, whereas explainability is restricted to post hoc analyses of model behavior.

Recent work at the intersection of mathematical optimization and interpretable machine learning demonstrates that fairness and explainability need not be post hoc corrections but can instead be embedded directly into model training. Carrizosa et al. (2025) propose a mixed-integer linear optimization formulation that trains tree ensembles, simultaneously trading off classification accuracy, sparsity, and group fairness with the formulation scaling linearly in the number of observations. Complementing this, Röber et al. (2025) introduce a column-generation approach for rule-based classification that is scalable to large data sets, provably handles fairness constraints, and returns optimal rule weights that directly indicate feature importance. Together, these contributions establish a principled paradigm in which interpretability, fairness, and predictive performance are treated as coequal objectives within a unified optimization framework.

The literature on explainable AI (XAI) responded with a rich ecosystem of global and local explainability techniques—ranging from sensitivity analysis and partial dependence plots to local surrogate methods such as LIME and SHAP—that seek to clarify how features influence predictions. Yet, as this body of work makes clear, global explanations may obscure meaningful subgroup heterogeneity, whereas local explanations can be difficult to aggregate into coherent, actionable insights at the group or population

level. This unresolved gap is particularly consequential in socially impactful applications, in which understanding differential effects across subpopulations is central to fairness, accountability, and informed decision making.

The paper "Supervised Clustered Interpretability: Explainable Subgroup Discovery via Cluster Separation and Partial Dependence Disparity" directly addresses this gap by introducing clustered interpretability as a principled middle ground between global and local explanations. Rather than explaining individual predictions in isolation or averaging effects across an entire data set, the proposed framework learns subgroups whose feature–outcome relationships are internally coherent yet meaningfully distinct across clusters. The core contribution—the separation–disparity index—formalizes interpretability as an optimization objective that simultaneously rewards separation along prediction-relevant features and penalizes unnecessary heterogeneity in feature effects across clusters. By embedding this metric within a tailored particle swarm optimization procedure, the approach jointly learns clusters and their associated explanatory structures, effectively integrating interpretability into the modeling process rather than treating it as a post hoc diagnostic. Empirical results on synthetic data and a real-world Parkinson's disease case study demonstrate that this framework yields subgroup explanations that are both more faithful to underlying data-generating processes and more stable than those obtained from conventional clustering or standalone XAI methods.

In the context of this special issue, the paper exemplifies how methodological advances in computing can advance responsible AI and data science for social good. By explicitly targeting subgroup-level transparency, the clustered interpretability framework supports more equitable and context-sensitive use of predictive models, enabling practitioners to detect heterogeneous effects, avoid spurious conclusions driven by global averages, and design interventions tailored to distinct population segments. This contribution resonates strongly with the special issue's themes: it advances interpretability beyond explanation-as-justification toward explanation-as-understanding, aligns optimization objectives with ethical and societal considerations, and demonstrates how careful algorithmic design can enhance both technical rigor and social relevance. As such, the paper not only extends the theoretical landscape of explainable machine learning but also provides a practical and principled pathway for deploying AI systems that are more transparent, trustworthy, and impactful in socially consequential domains.

## 8. Concluding Remarks and Future Research Directions

In May 2023, when we launched the Special Issue on Responsible AI and Data Science for Social Good in the *INFORMS Journal on Computing*, the field of AI was at a pivotal moment as rapid advances in machine learning and data science were already transforming industries and public institutions. However, the broader societal impacts of these technologies were only beginning to receive serious and systematic attention from the computing research community. Only a few months before the launch of this special issue, generative AI systems such as ChatGPT entered the public sphere and dramatically accelerated global awareness of both the transformative potential and the societal risks of AI. Questions of fairness, accountability, transparency, robustness, and social impact moved from academic discussions to boardrooms, regulatory bodies, classrooms, and households worldwide.

Over the three years during which this special issue evolved, from its announcement in May 2023 to its completion in February 2026, the AI landscape underwent profound change. Generative models became embedded in workflows across sectors; policymakers introduced new regulatory frameworks; and public discourse increasingly centered on ethical, legal, and societal implications. Against this backdrop, the mission of this special issue became even more urgent and relevant.

The 13 papers selected in this special issue reflect both forward thinking and adaptability. They highlight how strong computational research can address real societal challenges that are also listed in the United Nations sustainable goals, also pushing methodological innovation forward. Together, they illustrate that responsible AI is not a peripheral concern but a central pillar of modern computing research.

This paradigm shift from pregenerative to postgenerative AI requires a reorientation of research priorities toward developing new models for "Responsible AI and Data Science for Social Good." These models now need to be understood as a high-dimensional, multiobjective optimization challenge operating within dynamic sociotechnical ecosystems. It requires new computational methods capable of modeling systemic risk, long-term societal externalities, and human–AI collaboration at scale. Table 1 outlines key future research directions by contrasting pregenerative AI priorities with emerging postgenerative challenges and identifying the computational and optimization frameworks needed to responsibly utilize generative AI for societal benefit.

**Table 1.** Responsible AI Research Themes: Pregenerative and Postgenerative AI Shift

| Research theme | Pregenerative AI focus | Postgenerative AI shift | Key research questions (postgenerative AI) | Methodological approaches |
|---|---|---|---|---|
| Fairness and bias | Bias detection in predictive models (classification, risk scoring) | Bias propagation in foundation models across multiple modalities | • What computation methods can be used to quantify how bias propagates through generative AI pipelines?<br>• How can fairness constraints be embedded directly into generative decoding?<br>• How can multiobjective optimization in generative AI balance accuracy, diversity, and equity simultaneously? | **Computational methods**: causal inference models; adversarial robustness modeling; algorithmic auditing systems<br>**Optimization**: multiobjective optimization; constrained nonlinear optimization |
| Explainability and transparency | Estimate feature importance and contribution to model outputs; improve transparency, trust, and regulatory compliance in traditional machine learning systems | Explaining probabilistic reasoning and emergent behavior in large language models | • How can internal representation learning in foundation models be computationally mapped?<br>• How can explanation fidelity be optimized in generative AI systems, balancing other objectives such as accuracy and fairness?<br>• How do we optimize interpretability–performance trade-offs? | **Computational methods**: mechanistic interpretability; probing methods; counterfactual simulation models<br>**Optimization**: sparse optimization; bilevel optimization |
| Governance and accountability | Document machine learning models and data sets for transparency; use model cards for model details and intended use; use data sheets to describe data set characteristics | Life cycle governance across large generative and multimodal models' providers and downstream deployers and coordinate responsibilities between model creators, fine-tuners, and deployers | • How can accountability be formalized across multiactor generative AI ecosystems?<br>• Can regulatory compliance for generative AI be dynamically optimized under frameworks such as the EU AI Act?<br>• How can we computationally model systemic risk accumulation across interconnected generative AI models, providers, and applications? | **Computational methods**: traceability graph architectures; blockchain-based audit trails<br>**Optimization**: dynamic programming; robust optimization |
| Misinformation and synthetic media | Detection of bots, fake news, misinformation, and disinformation | Detection and mitigation of deepfakes and generative AI–generated persuasion at scale and address multimodal AI outputs (text, image, audio, video) | • How can adversarial attacks and manipulative strategies in generative AI systems be anticipated and countered computationally?<br>• How can detection systems for generative AI content be optimized to minimize both false positives and false negatives?<br>• What scalable architectures and frameworks can reliably verify the provenance of generative and multimodal AI outputs? | **Computational methods**: graph neural networks; adversarial learning systems<br>**Optimization:** min–max optimization; game-theoretic optimization |
| Privacy and security | Differential privacy and federated learning | Memorization risks, prompt injection, retrieval-augmented vulnerabilities | • How can memorization and data leakage risks in generative and multimodal foundation models be computationally detected and mitigated?<br>• How can constrained optimization be used to embed privacy guarantees in generative and multimodal AI models during training?<br>• How do we optimize privacy–utility trade-offs in generative and multimodal AI systems? | **Computational methods**: differential privacy and the Laplace mechanism; secure multiparty computation for federated learning<br>**Optimization:** convex optimization; Lagrangian relaxation |

**Table 1.** (Continued)

| Research theme | Pregenerative AI focus | Postgenerative AI shift | Key research questions (postgenerative AI) | Methodological approaches |
|---|---|---|---|---|
| Human–AI collaboration | Decision support systems | Semiautonomous AI copilots and agentic systems | • How can task allocation between humans and generative agents be optimized for social welfare?<br>• How can automation bias and overreliance on generative AI systems be computationally modeled?<br>• Can autonomy levels in generative AI systems be dynamically optimized for performance, safety, and accountability? | **Computational methods**: reinforcement learning from human feedback; agent-based modeling; human-in-the-loop simulation systems<br>**Optimization**: stochastic optimization; Markov decision processes |
| Environmental sustainability | Energy efficiency of machine learning pipelines | Carbon-intensive training and inference of large generative and multimodal AI systems | • How can energy-efficient architectures be computationally designed and optimized for large generative and multimodal AI systems?<br>• How can inference workloads across generative and multimodal AI systems be optimized to minimize carbon emissions through smart routing and scheduling?<br>• How can generative and multimodal AI systems be designed to maximize social impact, minimizing computational cost? | **Computational methods**: energy profiling algorithms; neural architecture search; carbon accounting models<br>**Optimization**: integer programming; resource allocation optimization |
| Inclusive and ethical AI | AI access for underserved populations; ethical AI principles and guidelines | Generative AI as accessibility amplifier or digital divider, and operationalizing pluralistic value alignment in foundation models to ensure ethical generative AI | • Can equitable performance in generative AI systems be achieved through fairness-constrained optimization?<br>• How can generative and multimodal AI systems allocate and optimize resources to maximize positive impact for marginalized communities?<br>• How can optimization frameworks reconcile conflicting ethical objectives across generative and multimodal AI systems? | **Computational methods**: preference learning models; multiagent aggregation architecture<br>**Optimization**: multiobjective evolutionary optimization; resource allocation models |
| Social impact measurement | Accuracy and performance metrics | Long-term societal externalities and systemic welfare effects | • What novel algorithmic approaches can estimate and mitigate the long-term societal impacts of generative and multimodal AI systems?<br>• Can welfare maximization in generative AI systems be formulated as a multiobjective optimization problem balancing social impact, fairness, and computational efficiency?<br>• How can we computationally optimize interventions in generative and multimodal AI systems under uncertainty, balancing effectiveness, fairness, and risk? | **Computational methods**: system dynamics modeling; computational social simulation; causal graphical models<br>**Optimization**: multicriteria decision optimization; robust stochastic optimization |

## References

Bakshi N, Kim J, Randhawa RS (2025) Service operations for justice-on-time: A data-driven queueing approach. *Manufacturing Service Oper. Management* 27(1):305–321.

Baltrušaitis T, Ahuja C, Morency LP (2018) Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Machine Intelligence* 41(2):423–443.

Barocas S, Hardt M, Narayanan A (2023) *Fairness and Machine Learning: Limitations and Opportunities* (MIT Press, Cambridge, MA), 1–294.

Bawack R, Dennehy D, Kumi CA, Boutchouang W (2025) AI analytics in enhancing patient-centered care through wearables: A cross-country analysis. *Inform. Systems Frontiers* 27(6):2631–2649.

Blumstein A (2007) An OR missionary's visits to the criminal justice system. *Oper. Res.* 55(1):14–23.

Brooks JP (2012) The court of appeals of Virginia uses integer programming and cloud computing to schedule sessions. *Interfaces* 42(6):544–553.

Carrizosa E, Kurishchenko K, Romero Morales D (2025) On enhancing the explainability and fairness of tree ensembles. *Eur. J. Oper. Res.* 323(2):599–608.

Chohlas-Wood A, Levine E (2019) A recommendation engine to aid in identifying crime patterns. *INFORMS J. Appl. Anal.* 49(2):154–166.

Delen D (2020) *Predictive Analytics: Data Mining, Machine Learning and Data Science for Practitioners* (Pearson FT Press, Hoboken, NJ).

Dimas GL, Konrad RA, Lee Maass K, Trapp AC (2022) Operations research and analytics to combat human trafficking: A systematic review of academic literature. *PLoS One* 17(8):e0273708.

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. *Proc. 3rd Innovations Theoret. Comput. Sci. Conf.* (ACM, New York), 214–226.

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, et al. (2018) AI4people—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Machines* 28(4):689–707.

Freeman N, Keskin BB, Bott GJ (2022) Collaborating with local and federal law enforcement for disrupting sex trafficking networks. *INFORMS J. Appl. Anal.* 52(5):446–459.

Freeman N, Bott G, Keskin B, Parton J, Cochran J (2026) Linking multisite sex ad data at the individual level to aid counter-trafficking efforts. *Manufacturing Service Oper. Management* 28(1):133–152.

Gillespie T (2018) *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press).

Glikson E, Woolley AW (2020) Human trust in artificial intelligence: Review of empirical research. *Acad. Management Ann.* 14(2):627–660.

Gorwa R, Binns R, Katzenbach C (2020) Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data Soc.* 7(1):1–15.

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. *Adv. Neural Inform. Processing Systems*, vol. 29.

Informs (2025) Court backlogs are clogging the system—New research finds a surprising fix. https://www.informs.org/News-Room/INFORMS-Releases/News-Releases/Court-Backlogs-Are-Clogging-the-System-New-Research-Finds-a-Surprising-Fix.

Jain M, Tsai J, Pita J, Kiekintveld C, Rathi S, Tambe M, Ordóñez F (2010) Software assistants for randomized patrol planning for the LAX airport police and the Federal Air Marshal Service. *Interfaces* 40(4):267–290.

Konrad RA, Maass KL, Dimas GL, Trapp AC (2023) Perspectives on how to conduct responsible anti-human trafficking research in operations and analytics. *Eur. J. Oper. Res.* 309(1):319–329.

Koski E, Das A, Hsueh PYS, Solomonides A, Joseph AL, Srivastava G, Johnson CE, et al. (2025) Towards responsible artificial intelligence in healthcare—Getting real about real-world data and evidence. *J. Amer. Medical Informatics Assoc.* 32(11):1746–1755.

Lazer DM, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, et al. (2018) The science of fake news. *Science* 359(6380):1094–1096.

Levine ES, Tisch J, Tasso A, Joy M (2017) The New York City Police Department's domain awareness system. *Interfaces* 47(1):70–84.

Lodi A, Sankaranarayanan S, Wang G (2024) A framework for fair decision-making over time with time-invariant utilities. *Eur. J. Oper. Res.* 319(2):456–467.

Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2017) Towards deep learning models resistant to adversarial attacks. Preprint, submitted June 19, https://arxiv.org/abs/1706.06083.

Memarian B, Doleck T (2023) Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. *Computers Ed. Artificial Intelligence* 5:100152.

Miao F, Holmes W, Huang R, Zhang H (2021) *AI and Education: A Guidance for Policymakers* (UNESCO Publishing, Paris).

Porayska-Pomsta K, Holmes W, Nemorin S (2023) The ethics of AI in education. du Boulay B, Mitrovic A, Yacef K, eds. *Handbook of Artificial Intelligence in Education* (Edward Elgar Publishing, Cheltenham, UK), 571–604.

Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, eds. (2009) *Dataset Shift in Machine Learning* (MIT Press, Cambridge, MA).

Röber TE, Lumadjeng AC, Akyüz MH, Birbil Şİ (2025) Rule generation for classification: Scalability, interpretability, and fairness. *Comput. Oper. Res.* 183:107163.

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5):206–215.

Samorani M, Harris S, Blount LG, Lu H, Santoro MA (2022) Overbooked and overlooked: Machine learning and racial bias in medical appointment scheduling. *Manufacturing Service Oper. Management* 24(6):2825–2842.

Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. *Proc. Conf. Fairness Accountability Transparency* (ACM), 59–68.

Shahabsafa M, Terlaky T, Gudapati NVC, Sharma A, Wilson GR, Plebani LJ, Bucklen KB (2018) The inmate assignment and scheduling problem and its application in the Pennsylvania Department of Corrections. *Interfaces* 48(5):467–483.

Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter* 19(1):22–36.

Siemens G (2013) Learning analytics: The emergence of a discipline. *Amer. Behav. Sci.* 57(10):1380–1400.

Taruffo M (1998) Judicial decisions and artificial intelligence. *Judicial Applications of Artificial Intelligence* (Springer), 207–220.

Trocin C, Mikalef P, Papamitsiou Z, Conboy K (2023) Responsible AI for digital health: A synthesis and a research agenda. *Inform. Systems Frontiers* 25(6):2139–2157.

Van Dijck J, Poell T, De Waal M (2018) *The Platform Society: Public Values in a Connective World* (Oxford University Press).

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. *Science* 359(6380):1146–1151.