Journal of Data Science 24 (1), 1–3
January 2026
DOI: 10.6339/26-JDS241EDI
Editorial

Editorial: Statistical Aspects of Trustworthy Machine Learning

Stephanie C. Hicks[1,*], Keegan Korthauer[2], Xiaotong Shen[3], Jun Yan[4], and Hao Helen Zhang[5]

[1]*Department of Biostatistics, Johns Hopkins University, USA*
[2]*Department of Statistics, University of British Columbia, Canada*
[3]*Department of Statistics, University of Minnesota, USA*
[4]*Department of Statistics, University of Connecticut, USA*
[5]*Department of Mathematics, University of Arizona, USA*

Machine learning methods are now routinely embedded in scientific discovery, healthcare, public policy, education, and many other domains where decisions carry substantial consequences. Although methodological advances have led to remarkable gains in predictive performance, concerns about trustworthiness have become increasingly prominent. Issues of interpretability, fairness and transparency, privacy preservation, and robustness shape whether machine learning systems are scrutinized, accepted, and relied upon in practice. These concerns are inherently statistical, involving modeling assumptions, uncertainty quantification, validation, and the interaction between data, algorithms, and human judgment. This special issue of the *Journal of Data Science* focuses on statistical aspects of trustworthy machine learning and brings together contributions that examine these issues across foundations, methodology, computation, applications, education, and community discussion.

The immediate impetus for this special issue was the Banff International Research Station workshop on "Statistical Aspects of Trustworthy Machine Learning," held in February 2024. We, the guest editors of this special issue, also served as the organizers of the workshop, which was designed to place statistical reasoning at the center of discussions on trustworthy machine learning. The workshop emphasized focused, sustained discussion of core themes, including interpretability, fairness and transparency, robustness, and privacy, with particular attention to uncertainty, model checking, data limitations, and the ways in which methodological choices interact with human decision-making. A recurring theme was that many failures of trust arise not from a lack of algorithmic sophistication, but from insufficient attention to these statistical considerations. While the Banff workshop helped shape the questions and perspectives represented here, this special issue is not a proceedings volume. Several contributions were submitted by authors who did not participate in the workshop, reflecting broader engagement within the data science community with statistical approaches to trustworthiness.

The issue opens with a contribution in the *Philosophies of Data Science* section by Jiang et al. (2026), which differs from most contributions in a collection centered on machine learning. Rather than introducing a new algorithm, the authors articulate a "typicality principle" for inference, arguing that a model should be regarded as unwarranted when the observed data are sufficiently atypical under the posited theory. By bringing goodness-of-fit and model checking to the foreground of inferential reasoning, this paper provides a conceptual foundation for thinking about trustworthiness that extends beyond any specific modeling framework.

Several papers appear in the *Data Science Review* section and provide synthetic perspectives on key components of trustworthy machine learning. Focusing on interpretability and explain-

*Corresponding author. Email: shicks19@jhu.edu.

ability, Sankaran (2026) develop a principled vocabulary and connect modern interpretability techniques to classical statistical ideas such as parsimony and experimental design, with attention to how interpretability depends on audience goals. Causal reasoning is examined by Wang et al. (2026a), who compare the potential outcomes framework, structural equation models, and directed acyclic graphs, clarifying their relationships and respective strengths. Sequential decision-making is reviewed by Zhou (2026), who examine reinforcement learning from a statistical perspective and highlight connections to uncertainty quantification, causal inference, and dynamic treatment regimes. In a nontechnical contribution aimed at a broad readership, Fu (2026) provide an overview of AI for science and propose a three-phase framework to conceptualize AI's evolving role in scientific discovery, placing statistical data science within wider scientific and societal contexts.

Methodological developments are featured in the *Statistical Data Science* section. Addressing interpretability for functional inputs, Goode et al. (2026) propose an explainable machine learning pipeline that combines elastic functional principal component analysis with post hoc explanations that remain faithful to the original data structure. Reliability issues in large language models are examined by Song et al. (2026), who analyze the phenomenon of reward collapse in ranking-based alignment methods and introduce prompt-aware objectives that preserve meaningful variation in rewards. In neuroimaging applications, Liu et al. (2026) develop a subject-specific scalar-on-image regression model that captures individual heterogeneity through spatially structured sparsity, leading to gains in both interpretability and predictive performance. Privacy preservation is addressed by Yu et al. (2026), who introduce a differentially private Bayesian envelope regression framework based on sufficient statistic perturbation and show how formal privacy guarantees can be integrated with efficient Bayesian inference. Fairness considerations are examined by Uddin et al. (2026), who use simulation studies to study trade-offs between accuracy and fairness metrics such as statistical parity and equalized odds.

Computational aspects of trustworthy machine learning are represented in the *Computing in Data Science* section by Zhang et al. (2026). That paper addresses the challenges in deploying large pretrained transformer models by introducing a magnitude pruning approach based on a mixture Gaussian prior, with the goal of reducing model complexity while preserving predictive performance. Theoretical justification and empirical evaluation are combined to demonstrate how statistically motivated regularization can improve computational efficiency, stability, and deployability in large-scale models.

Trustworthiness also depends on how data science is practiced and communicated in applied settings. In the *Data Science in Action* section, D'Agostino McGowan et al. (2026) propose a framework for quantifying the alignment between a data analyst and an intended audience. By formalizing alignment as the degree to which analytic principles and expectations are shared, the paper frames trustworthiness as a property that emerges from the interaction between analysts, analyses, and audiences.

Issues of education and assessment are the focus of the Education in Data Science section. Wang et al. (2026b) examine challenges posed by AI-generated assignment submissions and document limitations of existing detection-based approaches. The paper is accompanied by discussion contributions and an authors' rejoinder, offering a range of perspectives on how educators are responding. Beyond concerns about academic integrity, the discussion reflects a growing recognition that AI is becoming embedded in academic and professional workflows. Contributors describe concrete responses, including redesigning assessments to foreground reasoning and decision-making, increasing transparency around acceptable AI use, and incorporating AI tools into coursework as objects of instruction rather than technologies to be excluded. Together, these

contributions frame trustworthiness in education as an evolving practice shaped by pedagogy, assessment design, and ongoing community debate.

This special issue reflects a collective effort by the statistical and data science communities to clarify the role of statistics in building trustworthy machine learning. We thank the participants of the Banff International Research Station workshop for the sustained discussions that helped shape the scope of this issue, the authors for their contributions and careful revisions, and the reviewers and discussants for their thoughtful evaluations and perspectives. The resulting collection highlights both the diversity of approaches and the shared concerns that arise when machine learning systems are developed and deployed in consequential settings, and it points toward continued opportunities for statistical thinking to inform trustworthy practice.

References

D'Agostino McGowan L, Peng RD, Hicks SC (2026). Quantifying the alignment of a data analysis between analyst and audience. *Journal of Data Science*, 24(1): 239–253. https://doi.org/10.6339/25-JDS1189

Fu V (2026). AI for science: Opportunities, challenges, and future directions. *Journal of Data Science*, 24(1): 106–124. https://doi.org/10.6339/25-JDS1214

Goode K, Tucker JD, Ries D, Hofmann H (2026). Explainable machine learning for functional data. *Journal of Data Science*, 24(1): 125–145. https://doi.org/10.6339/25-JDS1212

Jiang Y, Zhang Z, Martin R, Liu C (2026). The typicality principle and its implications for statistics and data science. *Journal of Data Science*, 24(1): 4–25. https://doi.org/10.6339/26-JDS1217

Liu LYf, Ma H, Liu Y, Zhu H (2026). Subject-specific scalar-on-image regression. *Journal of Data Science*, 24(1): 167–186. https://doi.org/10.6339/25-JDS1203

Sankaran K (2026). Data science principles for interpretable and explainable AI. *Journal of Data Science*, 24(1): 26–52. https://doi.org/10.6339/24-JDS1150

Song Z, Cai T, Lee JD, Su WJ (2026). Reward collapse in aligning large language models. *Journal of Data Science*, 24(1): 146–166. https://doi.org/10.6339/25-JDS1201

Uddin MB, Yin M, Dasgupta N (2026). A designed look at artificial intelligence from the lens of fairness. *Journal of Data Science*, 24(1): 203–217. https://doi.org/10.6339/26-JDS1219

Wang L, Richardson T, Robins J (2026a). Causal inference: A tale of three frameworks. *Journal of Data Science*, 24(1): 53–85. https://doi.org/10.6339/25-JDS1211

Wang S, Xu L, Liu J, Zhai Y (2026b). Addressing the challenges of AI-generated assignment submissions in education: Insights and strategies. *Journal of Data Science*, 24(1): 254–260. https://doi.org/10.6339/25-JDS1208

Yu P, Jiang Y, Su Z, Wu J, Kang L, Jiang B (2026). Differentially private Bayesian envelope regression via sufficient statistic perturbation. *Journal of Data Science*, 24(1): 187–202. https://doi.org/10.6339/25-JDS1194

Zhang M, Sun Y, Liang F (2026). Magnitude pruning of large pretrained transformer models with a mixture Gaussian prior. *Journal of Data Science*, 24(1): 218–238. https://doi.org/10.6339/24-JDS1156

Zhou Y (2026). Reinforcement learning: A statistical perspective. *Journal of Data Science*, 24(1): 86–105. https://doi.org/10.6339/25-JDS1205