frontiers
in Big Data

EDITORIAL
published: 12 August 2021
doi: 10.3389/fdata.2021.742589

Editorial: Ethical Machine Learning and Artificial Intelligence

Novi Quadrianto[1,2]*, Björn W. Schuller*[3,4]** and Finnian Rachel Lattimore*[5]

[1]*PAL – Predictive Analytics Lab, University of Sussex, Brighton, United Kingdom,* [2]*BCAM Severo Ochoa Strategic Lab on Trustworthy Machine Learning, Bilbao, Spain,* [3]*GLAM – Group on Language, Audio, & Music, Imperial College London, London, United Kingdom,* [4]*EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany,* [5]*Gradient Institute, Sydney, NSW, Australia*

Keywords: ethics, artificial intelligence, machine learning, fairness, accountability, transparency, trustworthiness, General Data Protection Regulation

Editorial on the Research Topic

Ethical Machine Learning and Artificial Intelligence

1 INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) have increasingly become an every-day reality for most of us (Elliott, 2019). Typical algorithmic assessment methods, used for predicting human outcomes such as recruitment, bail decisions, mortgage approvals, and insurance premiums, among many others, are currently being trialled and subsequently deployed. Hence, the ethical and legal requirements are moving into the foreground when developing novel AI and machine learning algorithms (Hagendorff, 2020). For example, the United States' Fair Credit Reporting Act and European Union's General Data Protection Regulation (GDPR) prescribe that data must be processed in a way that is fair/unbiased—a challenge for AI (Mehrabi et al., 2019). GDPR also alludes to the right of an individual to receive an explanation about decisions made by an automated system such as by explainable AI (XAI) (Gunning et al., 2019).

Here, based on a recent research topic held in Frontiers in Big Data, we provide an overview on the authors' views and contributions.

This research topic covers but is not limited to the fields of fairness, accountability, transparency, and trustworthiness (Baird et al., 2019), and covers methods such as causality and counterfactual reasoning, reinforcement learning, and probabilistic approaches.

2 LITERATURE REVIEW

The research topic provides two overviews on the field.

In the first, Wells and Bednarz discuss in "*Explainable AI and Reinforcement Learning – A Systematic Review of Current Approaches and Trends*" 25 studies selected from 520 search hits on this recent topic. Thereby, they focus on "visualisation, query-based explanations, policy summarisation, human-in-the-loop collaboration, and verification" which they identify as trends. As others, they name the urge for user evaluations including laymen of explanations and find examples often over-

OPEN ACCESS

Edited and reviewed by:
Sriraam Natarajan,
The University of Texas at Dallas,
United States

****Correspondence:***
Björn W. Schuller
schuller@ieee.org

Specialty section:
This article was submitted to
Machine Learning and Artificial
Intelligence,
a section of the journal
Frontiers in Big Data

Received: *16 July 2021*
Accepted: *22 July 2021*
Published: *12 August 2021*

Citation:
Quadrianto N, Schuller BW and
Lattimore FR (2021) Editorial: Ethical
Machine Learning and
Artificial Intelligence.
Front. Big Data 4:742589.
doi: 10.3389/fdata.2021.742589

simplified going hand-in-hand with lack in scalability, while provision of comprehensible explanations remains a key challenge. Further, they consider more progressive visualisation approaches under-exploited including multimodal and immersive forms of visualisation. Ideally, in the authors' opinion, such would be combined with "well articulated explanations".

Next, in their mini review "*Considerations for a More Ethical Approach to Data in AI: On Data Representation and Infrastructure*", Baird and Schuller observe that data infrastructures are increasingly managed more democratically, as decentralisation fosters transparency and therefore can help better cope with selection-bias. Their review deals with AI-targeted data representation and infrastructures focussing on "auditing, benchmarking, confidence and trust, explainability and interpretability" as key aspects that require attention—ideally also in an interdisciplinary endeavour. As to auditing, in multimodal applications, the authors require standards per modality to lead to accurate benchmarking. Further, they support the view that confidence and trust are benefited by "diverse representations of human data"—the latter also boosting explainability to all users given "inherent human-like attributes". The authors attest energy put into these aspects by the community, but in particular demand for increased standardisation.

3 TECHNICAL APPROACHES

The research topic further includes three technical solutions.

First, in "*The Moral Choice Machine*", Schramowski et al. demonstrate that one can "extract deontological ethical reasoning" with machine learning from human written texts concerning right or wrong conduct. The authors provide prompts and responses and define a bias score based on the score of positive and negative responses. Likewise, they reach to the Moral Choice Machine (MCM), that determines this score per sentence applying Universal Sentence Encoder embeddings to cater for context. By that, they observe that textual databases bear "recoverable and accurate imprints of our social, ethical and moral choices". Further, picking selected databases from different epochs, they find reflection on the evolution of these aspects. Similarly, the authors consider different cultural sources. Ultimately, this leads to their view that "moral biases can be extracted, quantified, tracked, and compared across cultures and over time". As future work, the authors name the possibility to alter the embeddings in targeted ways, such as to eliminate gender stereotypes. They further suggest having the moral choice machine in interactive robots enabled with active learning to have users correct potential biases. Finally, they suggest targeted alteration of the text sources for observation of effects.

In "*Tuning Fairness by Balancing Target Labels*", Kehrenberg et al. deal with bias in the output as challenge. To this end, they add a latent target output to cater for a unified approach, apply marginalisation rather than constraints problem, and provide for a possibility to integrate knowledge on target unbiased outputs. The authors argue that fairness is usually mainly handled by statistical (group) or individual notions and belief that both are needed for algorithmic fairness. Their approach can be learnt from an implicitly balanced corpus, hence enabling demographic parity and equality of opportunity. They also indicate avenues towards an extension aiming at conditional demographic parity as well. Finally, their general approach uniquely provides for a target rate to control the realisation of the fairness constraint. However, it will need extensions for predictive parity group or individual fairness.

As a third example of algorithmic contribution to a more ethical approach serve Ramanan and Natarajan's with "*Causal Learning From Predictive Modeling for Observational Data*". They apply causal Bayesian networks to model causal relationships between data-learnt model variables sequentially using context-specific and mutual independence. Likewise, potential causal relationships are first found. Subsequentially, their strength is determined. The authors verify this approach on benchmark networks and find superiority over current alternatives.

4 DISCUSSION

Card and Smith finally discuss "*On Consequentialism and Fairness*", focusing on the outcome. They argue that consequentialism has its deficits such as lacking in an amenable choice of actions, but is a suited mean to highlight issues in AI fairness such as "who counts", disadvantages of policy application, or the relative weight of the future. The authors give a consequentialism-based critique of prevailing fairness definitions in AI. They further also take an AI viewpoint on consequentialism. Finally, they elaborate on learning and randomisation in the context of AI ethics.

5 CONCLUSION

As all authors highlight, a more ethical approach is needed to data in AI. However, algorithmic solutions can be and were partially given also here. Accordingly, there is a call to action also for those providing AI algorithms in the first place to actively work on solutions to benefit and protect all users of AI and society.

AUTHOR CONTRIBUTIONS

BS wrote the manuscript. NQ edited it. All authors led the underlying research topic.

ACKNOWLEDGMENTS

We express our sincere gratitude to all authors and reviewers that helped putting together the research topic.

REFERENCES

Baird, A., Hantke, S., and Schuller, B. (2019). Responsible and Representative Multimodal Data Acquisition and Analysis: on Auditability, Benchmarking, Confidence, Data-reliance & Explainability. *Clin. Orthop. Relat. Res.* arXiv [preprint] arXiv: 1903.07171. Available at: http://arxiv.org/abs/1903.07171/.

Elliott, A. (2019). *The Culture of AI: Everyday Life and the Digital Revolution*. London: Routledge.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G. Z. (2019). XAI-explainable Artificial Intelligence. *Sci. Robot.* 4, eaay7120. doi:10.1126/scirobotics.aay7120

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines.* 30, 99–120. doi:10.1007/s11023-020-09517-8

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *Clin. Orthop. Relat. Res.* arXiv [preprint] arXiv: 1908.09635. Available at: http://arxiv.org/abs/1908.09635/.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.