Preface: AI Ethics

Artificial intelligence (AI) is no longer just an academic endeavor. Today, AI is helping us run nearly every part of our infrastructure and our lives, whether we consider health care, transportation, finance, commerce, agriculture, criminal justice, or human capital management. As an emerging technology so intertwined with society, and with such profound consequences, it is essential that AI be safe, fair, ethical, and trustworthy.

This special issue of the *IBM Journal of Research and Development* explores various facets of AI ethics, starting from the problem of value alignment, which is the task of eliciting and representing the values of society—what is considered good and bad—so that the goals and behaviors of AI systems can be designed to follow them. One particularly thorny concept to elicit from society is fairness, which can be defined in innumerable ways even in the restricted setting of allocation decisions. Once we have fairness values, aligning AI systems to those values may be accomplished via bias detection and mitigation algorithms. Fairness and many other values such as technical robustness and transparency may then be tested and governed, including in specific application domains.

The first two papers investigate methods for value alignment. The first paper, by Balakrishnan et al., focuses on learning ethical priorities for online AI systems. The authors propose a multi-armed bandit algorithm that incorporates context and behavioral constraints. They demonstrate the approach on movie recommendation and anticoagulant treatment decision examples.

The second paper, by Noothigattu et al., also utilizes contextual bandits, but for choosing between two policies over sequences of actions of a reinforcement learning agent: one that maximizes an environmental reward and one that has been learnt from humans with moral constraints. This approach is demonstrated on the videogame Pac-Man in which eating the ghosts is taken to be the immoral behavior.

The third paper presents a bias mitigation algorithm intended for pre-processing multimedia datasets, such as images, to ensure fair allocation decisions. Sattigeri et al. propose the Fairness GAN, a generative adversarial network that takes an existing dataset containing unwanted biases that yield systematic disadvantages to groups defined by protected attributes such as race and gender, and samples a new dataset that is similar but with much less systematic disadvantage.

In the fourth paper, Bellamy et al. introduce a new open-source Python toolkit, AI Fairness 360, which is a comprehensive collection of fairness metrics and state-of-the-art bias mitigation algorithms, as well as tutorials, interactive demos, and guidance materials. The toolkit is well-engineered and extensible, with the goal of translating methods from research labs to data scientists, data engineers, and developers deploying solutions in a variety of sectors.

The next set of papers transition to testing, reporting, and governing. The fifth paper, by Srivastava and Rossi, examines the problem of providing bias ratings to AI services, including composite services that involve several modules. They illustrate their approach on natural language processing tasks where gender bias may be an issue. For example, language translation from English to a middle language that lacks gender markings and back to English could reveal a bias.

The sixth paper, by Arnold et al., proposes FactSheets as a methodology to increase transparency of AI services. Suppliers of the technology would voluntarily release a declaration of conformity indicating the intended use of the service along with the results of several standardized tests conducted on the service, such as accuracy, fairness, and robustness.

In the seventh paper, Coates and Martin move beyond rating and reporting to governance. They propose a maturity framework that assesses an organization's capability to govern bias. Involving both technical and organizational aspects, the instrument they have developed could help AI development teams identify actions to improve the ethics of their outputs.

The final two papers are focused on specific application domains. Rodriguez et al., in the eighth paper, focus on social work, specifically child protective services. Performing predictive modeling on a dataset from the National Child Abuse and Neglect Data System, they investigate the choice of input features and choice of outcome variable on algorithmic fairness.

Simbeck, in the ninth paper, examines AI ethics in the context of human capital management functions such as predicting employee attrition and estimating employee skills. She applies ethical frameworks developed in other fields to human resource analytics, finding there to be five key principles: privacy, transparency, institutional review, opportunity to opt-out, and respect for the fact that people change over time, i.e., past behavior does not necessarily imply the same future behavior.

We are at an important juncture in human history: AI can augment our abilities for the good or for the bad, can behave in constructive or destructive ways, and can lead to a world full of equity or inequity. Technology is of course not the root of all ills nor the solution to them all, but it certainly has a role to play. The articles in this issue explore many sides of AI ethics; we hope you take their contents to heart and use them to help advance society in a positive direction.

Sameep Mehta
Senior Technical Staff Member
IBM Research, India

Francesca Rossi
Distinguished Research Staff Member
IBM Research, Thomas J. Watson Research Center

Kush R. Varshney
Principal Research Staff Member
IBM Research, Thomas J. Watson Research Center

Guest Editors

Digital Object Identifier: 10.1147/JRD.2019.2944775