Sociotechnical Perspectives on AI Ethics and Accountability

Nadin Kokciyan, *University of Edinburgh, Edinburgh, EH8 9AB, U.K.*
Biplav Srivastava and Michael N. Huhns, *University of South Carolina, Columbia, SC, 29208, USA*
Munindar P. Singh, *North Carolina State University, Raleigh, NC, 27695-8206, USA*

Suppose we were to develop a loan-processing tool based on artificial intelligence (AI) to process applications by people for financial loan products. The tool would consider application data and recommend whether to give a loan and for how much. It would even seek out prospective borrowers online for new business and offer loans. Or, suppose we were to develop a career coach that recommends career tracks and training based on a user's career goal, biosketch, and time and money available to invest in training.

Applications of AI in decision support are not hypothetical, and applications such as loan processing and career coaching are becoming mainstream. However, although like other algorithms, their inputs and outputs are data; these AI applications are embedded in society, their decisions and recommendations have direct effects on people's lives. Denial of a loan reduces financial options and may harm a borrower's wellbeing, while giving a loan but at usurious interest rates might expose a borrower to financial ruin. Likewise, whereas career advice can be valuable to someone who does not have strong mentors, narrow or biased career advice can impede their future and, through them, their family's prospects. The abovementioned AI applications illustrate settings where a person interacts with a single agent. Just as interesting, if not more, are applications in which AI can help multiple humans interact. For example, how nurses allocate resources in a hospital affects outcomes for patients. When the nurses are assisted by agents, would they be able to cooperate better and improve patient outcomes?

In general, biases and other misbehaviors of AI technologies are not always blatant or statistically observable. Therefore, we need regulatory structures to curtail such behaviors without sacrificing autonomy. Intuitively, the autonomy of an AI agent reflects the social autonomy of its primary stakeholder, and preserving autonomy provides a basis for flexibility by minimally constraining an agent's behavior. Accountability provides an important basis for organizing sociotechnical systems because it captures what the social entities in a sociotechnical system may expect from one another independent of their implementations.

It is, therefore, little wonder that the recent resurgence of AI applications has led to an increasing interest in ethics and accountability, as well as associated concerns such as fairness, liability, and verification. Accountability is multifaceted and can be viewed ethically (morally), financially, socially (fairness and diversity), and legally (according to both rule of law and rule by law). However, much of the literature on AI ethics takes an atomistic, single-agent perspective, such as the decision-making involved in the trolley problems, or the statistical aspects of machine learning algorithms, and addresses accountability simplistically in terms of tracing actions or imposing penalties for misbehavior.

In contrast, this special issue focuses on system-level perspectives on AI ethics and accountability, which consider how AI is embedded in technology while respecting the needs of society. Specific societal settings include a human or organization using an AI algorithm to arrive at decisions that affect others, as well as AI agents assisting humans in how they interact with each other and with existing social institutions. The articles we present cover diverse aspects of ethics and accountability from a sociotechnical perspective.

One of the challenges on AI ethics and accountability is to define these concepts formally. This step is necessary toward building computational reasoning frameworks to verify if some desired properties would hold under specified conditions. In "Reimagining Robust Distributed Systems Through Accountable MAS," Baldoni *et al.* propose a conceptual model to define accountability for designing robust multiagent systems. They base accountability on two dimensions: 1) the normative dimension—an agent can be asked about a task if only this is legitimate; and 2) the structural dimension—an agent should have control over a task to be accountable. This conceptual model provides guidance for organizations to enable accountability in their specific contexts. Baldoni *et al.* show how agents could gain robustness to perturbations by implementing such a model by extending a platform for developing agents to participate in a multiagent system.

In a multiagent setting, agents make collective decisions by using data obtained from various information

1089-7801
Digital Object Identifier 10.1109/MIC.2021.3117611
Date of current version 6 December 2021.

sources. In such a complex environment, it can be difficult to identify agents who are responsible for an action. This calls for the development of morally responsible agents, which is what Yazdanpanah *et al.* discuss in "Different Forms of Responsibility in Multiagent Systems: Sociotechnical Characteristics and Requirements." Specifically, they present sociotechnical characteristics of different forms of responsibility in multiagent systems, where agents could be responsible, blameworthy, accountable, or sanctionable based on the collective decision-making process involved. To make autonomous systems trustworthy, the agents should be empowered to reason about the forms of responsibility for addressing individual as well as societal needs. This work sets the stage for the development of reasoning frameworks to capture and reason about different forms of responsibility.

We cannot consider sociotechnical systems without their stakeholders. The next article in this issue helps close the loop by seeking to understand how users of AI tools perceive their trustworthiness in terms of ethics and accountability. In "Understanding User Perceptions of Trustworthiness in E-Recruitment Systems," Ogunniye *et al.* develop a prototype e-recruitment tool that helps its users understand how it ranks job applicants. They describe a user study to measure user perspectives on AI such as fairness, transparency, and trustworthiness. Their results suggest that users find AI tools fairer and more trustworthy when they are provided basic explanations about how they work.

Another article in this issue addresses accountability, though it was selected independently of this theme. In "Accountability as a Foundation for Requirements in Sociotechnical Systems," Chopra and Singh provide a metamodel for accountability based on concepts that go back to studies of norms in the law. They use this metamodel as a basis for capturing high-level requirements for how intelligent agents ought to interact with one another. Specifically, these requirements state what each party is accountable for, to whom, and under what circumstances. Doing so provides an implementation-independent way to characterize good behaviors in a sociotechnical system. Chopra and Singh demonstrate their approach on a case study of patient transfer between the emergency and clinical departments in a hospital.

We expect that these articles will jointly provide a valuable perspective in bringing trustworthy AI-based systems into practice. Arguably, the AI technology and the ecosystem in which the technology is developed and deployed should constitute sociotechnical systems that support accountability and thus engender trust—no less than traditional systems that exist for the same purpose but which do not rely upon AI technologies. Accountability can apply both to the outcome of the sociotechnical system and the process of reaching the outcome: Were the stakeholders' objectives achieved and were they achieved ethically? If it is wrong for a human loan officer to prey on the economically vulnerable, it should be wrong for an AI loan agent to prey on them and, moreover, the sociotechnical system should include mechanisms to regulate each AI agent's behavior and hold it to account. The sociotechnical systems we realize by bringing AI technologies into society should respect human values and engender trustworthy behavior.

NADIN KOKCIYAN is currently a Lecturer in Artificial Intelligence with the School of Informatics, University of Edinburgh, Edinburgh, U.K. Her primary focus is developing AI techniques to support decision-making in multiagent systems. She received the Ph.D. degree in computer engineering from Bogazici University, Istanbul, Turkey, in 2017. Contact her at nadin.kokciyan@ed.ac.uk.

BIPLAV SRIVASTAVA is currently a Professor of Computer Science with the AI Institute, University of South Carolina, Columbia, SC, USA. His research interests include neuro-symbolic methods for decision support, trusted AI, and sustainability. He received the Ph.D. degree in computer science from Arizona State University, Tempe, AZ, USA, in 2000. He is an ACM Distinguished Scientist, AAAI Senior Member, and IEEE Senior Member. Contact him at biplav.s@sc.edu.

MICHAEL N. HUHNS is currently the NCR Distinguished Professor Emeritus of Computer Science and Engineering at the University of South Carolina, Columbia, SC, USA. Prior to this, he was the Chair of the Department of Computer Science and Engineering. His research interests include multiagent systems, ontologies, and sociotechnical systems. He received the B.S. degree from the University of Michigan, Ann Arbor, MI, USA, in 1969, and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, CA, USA, in 1971 and 1975, respectively, all in electrical engineering. He is a Fellow of the IEEE, a Senior Member of the ACM, and a Fellow of the Association for the Advancement of Artificial Intelligence. Contact him at huhns@sc.edu.

MUNINDAR P. SINGH is currently a Professor in Computer Science and the Co-Director of the Science of Security Lablet, NC State University, Raleigh, NC, USA. His research interests include ethics and governance from a sociotechnical systems perspective. He is a Fellow of the IEEE, AAAI, and AAAS, and a Member of Academia Europaea. He received the B. Tech. degree in computer science and engineering from the Indian Institute of Technology, Delhi, India, in 1986 and the Ph.D. degree in computer sciences from the University of Texas, Austin, TX, USA, in 1993. He is a Former Editor-in-Chief of *IEEE Internet Computing* and *ACM Transactions on Internet Technology*. Contact him at singh@ncsu.edu.