Statement on Machine Learning* and Fairness

December 10, 2019
Japan Society for Artificial Intelligence, Ethics Committee
Japan Society for Software Science and Technology, Machine Learning Systems Engineering Group
IEICE, Information-Based Induction Sciences and Machine Learning Group

We, a group of researchers studying Machine Learning technologies and their applications (Japan Society for Artificial Intelligence, Ethics Committee; Japan Society for Software Science and Technology, Machine Learning Systems Engineering Group; and IEICE, Information-Based Induction Sciences and Machine Learning Group, hereinafter referred to as “we”) acknowledge that Machine Learning may interact with concepts of fairness in a way that is problematic. We would like to share our thoughts on how we believe the issue should be addressed and make the following two important points:

(1) Machine learning is nothing more than a tool to assist human decision making, and
(2) We are committed to improving fairness in society by studying possible uses of Machine Learning

Background

We understand that there is growing concern that the improper use of Machine Learning may have a negative impact on the fairness of outcomes. For example, in October 2018, Reuters reported that Amazon noticed that the Machine Learning system used in their hiring process was resulting in decisions that showed bias against women, and Amazon stopped using the system [1]. More generally, we recognize that improper use of Machine Learning may, intentionally or unintentionally, affect the fairness of outcomes in various contexts (see [2]).

Machine Learning is nothing more than a tool

Machine learning is a tool and human beings decide whether and how to use it. Machine learning has the potential to make a significant contribution to the prosperity of society, but if used inappropriately, it may also cause harm to society. To the extent that Machine Learning predicts the future based on past examples, the future predicted based on a biased past may carry that bias forward. If we want a better future than the biased past, humans may need to carefully intervene in the Machine Learning process to ensure that outcomes are fair.

At the same time, the contours of “what is fair” are determined by society, and advancements in and deployment of science and technology need to be consonant with society’s values. In order to make proper use of this Machine Learning tool, we must understand exactly how it interacts with our society’s values of “fairness”, evaluate its risks, and agree on how to implement countermeasures to deal with the identified and realised risks. This needs to be understood and dealt with not only by us Machine Learning researchers but also by engineers, end users, managers, organizations, and society as a whole.

We contribute to fairness by aligning Machine Learning

We are committed to avoiding the risks of improper use of Machine Learning and striving to solve the problems, from both code of conduct and technology development points of view. Recently, the IEEE Global Initiative published Ethically Aligned Design, First Edition [3] in which the misuse of Machine Learning is prohibited and specific countermeasures are shown. In Japan, the Japanese Society for Artificial Intelligence defined its Ethical Guidelines in 2017 to serve as a moral foundation for its members as well as to increase their awareness of their social responsibilities and to encourage effective communications with society [4]. Together with various stakeholders in Japan, we discussed how advanced information technology should be used in society, and the results of these discussions were published in March 2019 as Social Principles of Human-Centric AI [5]. One guiding principle of this work is ‘diversity and inclusion’. Also, it is clearly stated that stakeholders should be responsible for fair decision making and accountable for outcomes when advanced information technology is used.

In light of this, we are undertaking research on how to evaluate quantitatively and realize various aspects of fairness. Fairness in Machine Learning has become a prominent topic in recent symposia, and the number of research papers on fairness is increasing worldwide. When mathematically analyzing the various concepts of fairness using Machine Learning terms, it can be seen that there are many variations of fairness. As such, the concept of “fairness” can be made clearer by re-expressing various criteria in terms of Machine Learning. Using this approach, we hope not only to prevent undesired outcomes when using Machine Learning, but also to promote discussions regarding the various definitions of fairness.

Looking ahead

The above two points led us to think about what to do next. The issue of fairness needs to be discussed in an ongoing manner from the perspective of both what technology can do and what society wants. As society's interest in fairness in Machine Learning increases, we should be more sensitive to our social responsibilities and promote open dialogue with everybody in our society.

* Systems using Machine Learning technology are sometimes referred to as "artificial intelligence". However, the expression "artificial intelligence" can also refer to a prospective technology or system that may or may not be invented in the future as the outcome of artificial intelligence research. This statement is specifically concerned with existing “Machine Learning” technology, not speculative technology.

References

[1] Amazon scraps secret AI recruiting tool that showed bias against women., https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[2] O'Neil, Cathy. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, 2016. Cathy O'Neill (Author), Naoko Kubo (Translation) ", 2018.

[3] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (The IEEE Global Initiative), Ethically Aligned Design-A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition, https://ethicsinaction. ieee.org/, 2019.

[4] Japanese Society for Artificial Intelligence, Ethical Guidelines, http://ai-elsi.org/archives/514

[5] Cabinet Office, Social Principles of Human-Centric AI, https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf.