Conceptualizing fairness and trust in algorithmic decision-making

Min Kyung Lee, Carnegie Mellon University (mklee@cmu.edu, www.cs.cmu.edu/~mklee)

Algorithms increasingly govern and manage many functions of society. More than ever it is critical to ensure that these algorithms make fair and trustworthy decisions. I argue that we should go beyond examining the computational performances of algorithms and consider the important, yet less acknowledged contexts in which algorithmic decisions are embedded. In this whitepaper, I lay out initial contextual factors that shape the ways we can conceptualize the fairness and trustworthiness of algorithmic decisions.

Social context: Stakeholders in creation and usage of algorithmic decisions

We first need to identify the different roles that people play in creating and using algorithmic decisions, because the meanings of algorithmic fairness and trust vary depending on which stakeholders' perspectives we consider.

Algorithm developers. Developers directly define algorithms' inputs and outputs, performance metrics, and the rules and factors that algorithms use to make the decisions. Developers' beliefs and development practices can intentionally or unintentionally bias algorithms' decisions (Barocas & Selbst, 2016).

Users of algorithmic decisions. Users utilize algorithmic decisions as a tool. One category of users is decision-makers who execute tasks based on algorithmic decisions for work. Examples include managing factory automation, diagnosing and recommending medical conditions and treatments, creating a financial investment portfolio, patrolling based on predictive policing, and determining criminal sentencing. A long line of research on decision-aids is pertinent to this context. Another category of user is the information consumer, who uses algorithmically curated content and recommendations, such as search results, social media sites, and online news. Users in this context have less control over the algorithms than developers, and the mechanisms that underlie algorithmic decisions are mostly hidden (e.g., Rader & Gray, 2015); but they can decide whether to use algorithmic decisions or not, and often can give feedback on or customize the algorithms.

People affected by algorithmic decisions. Another stakeholder is people who are affected by algorithmic decisions. Examples include workers in on-demand work platforms, citizens in neighborhoods where predictive policing is applied, job seekers whose applications are reviewed by algorithms, and students in online education platforms whose exams are

Figure 1. Different stakeholders in creation and usage of algorithmic decisions

graded by algorithms. In this situation, people often cannot refute the decisions (Lee et al., 2015), and have little to no control over the algorithms. This is an emerging category of users, who deserve attention because of the new roles that algorithms take in social functions. There is relatively little research in this area.

Task context: People's beliefs about human versus algorithmic skills
Another factor is task characteristic–whether people think a certain tasks can be done better by humans than algorithms or that algorithms can do as well as or better than humans. Despite recent advances in machine learning and artificial intelligence, which make algorithms capable of executing tasks that people previously have believed only humans can do (such as speech recognition), people still believe algorithms and machines cannot make subjective judgments and process emotions (Gray, Gray and Wegner, 2007; Waytz and Norton, 2014). Lee's work (2017) suggests that people's trust and fairness perceptions of algorithmic decisions are influenced by these beliefs. For work assignment and scheduling tasks, both algorithmic and human decisions were perceived to be equally fair and trustworthy. On the other hand, for hiring and evaluating tasks, algorithmic decisions were perceived as significantly less fair and trustworthy than human decisions.

Value context: Social construction of algorithmic fairness and trustworthiness
The final factor is the values that different social groups hold, which influence their concepts of fairness and trustworthiness. This context-dependent nature makes it difficult to define algorithmic fairness. For example, Lee and Baykal (2017)'s work shows the gaps between mathematically proven fair division algorithms and varying fairness notions that different social groups held, which led to unfair perceptions of algorithmic decisions. In the context of a donation allocation algorithm, different stakeholders held varying concepts of fair donation distribution, which made it challenging to operationalize fairness in algorithms (Lee, Kim, Lizarondo, 2017).

What trustworthiness means also depends on context. In the case of users, especially decision makers, trust includes people's attitudes toward algorithmic accuracy, performance and levels of control. On the other hand, in the case of people affected by algorithmic decisions, trustworthiness also depends on perceived fairness (whether they think that algorithmic decisions are made fairly for all), and whether the organization employs a method for them to refute the decisions.

Going forward. In the above section, I illustrate initial contextual factors for conceptualizing algorithmic fairness and trustworthiness. In the workshop, I hope to discuss with participants how we can build systematic understanding of algorithmic fairness and trustworthiness accounting for these contextual factors.

References
Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review.

Gray H.M., Gray, K. & Wegner, D.M. (2007). Dimensions of mind perception. *Science 315*(5812), 619-619.

Lee, M. K. (2017). Understanding perception of algorithmic decisions: Fairness, trust and emotion in response to algorithmic management. Conditionally accepted to *Big Data & Society*.

Lee, M. K., Kim, J. & Lizarondo, L. (2017). A human-centered approach to algorithmic services: Considerations for fair and motivating smart community service management that allocates donations to non-pro t organizations. In *Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems (CHI 2017)*, 3365-3376.

Lee, M. K. & Baykal S. (2017) Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW 2017)*, 1035-1048.

Lee, M. K., Kusbit, D., Metsky, E. & Dabbish, L. (2015). Working with machines: The impact of algorithmic, data-driven management on human workers. In *Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems (CHI 2015)*, 1603-1612.

Rader, E., & Gray, R. (2015). Understanding user beliefs about algorithmic curation in the Facebook news feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI 2015)*, 173-182.

Waytz, A. & Norton, M.I. (2014). Botsourcing and outsourcing: Robot, British, Chinese, and German workers are for thinking—not feeling—jobs. *Emotion, 14*(2), 434-444.