Decision systems that respect privacy, fairness

August 11 2017, by Vidya Palepu

Increasingly, decisions and actions affecting people's lives are determined by automated systems processing personal data. Excitement about these systems has been accompanied by serious concerns about their opacity and threats they pose to privacy, fairness, and other values. Examples abound in real-world systems: Target's use of predicted pregnancy status for marketing; Google's use of health-related search queries for targeted advertising; race being associated with automated predictions of recidivism; gender affecting displayed job-related ads; race affecting displayed search ads; Boston's Street Bump app focusing pothole repair on affluent neighborhoods; Amazon's same day delivery being unavailable in black neighborhoods; and Facebook showing either "white" or "black" movie trailers based upon "ethnic affiliation."

Recognizing these concerns, CyLab's Anupam Datta, associate professor of electrical and computer engineering at Carnegie Mellon's Silicon Valley campus, will lead a $3 million National Science Foundation project on accountable decision systems that respect privacy and fairness expectations. The project seeks to make real-world automated decision-making systems accountable for privacy and fairness by enabling them to detect and explain violations of these values. The project will explore applications in online advertising, healthcare, and criminal justice, in collaboration with domain experts.

The project team includes Matthew Fredrikson, assistant professor of computer science, and Ole Mengshoel, principal systems scientist in electrical and computer engineering. The project also marks a

collaboration between CMU, Cornell Tech, and the International Computer Science Institute; additional contributors are Helen Nissenbaum, professor of information science at Cornell, Thomas Ristenpart, associate professor of computer science at Cornell, and Michael C. Tschantz, senior researcher at the International Computer Science Institute in Berkeley.

"A key innovation of the project is to automatically account for why an automated system with artificial intelligence components exhibits behavior that is problematic for privacy or fairness," says Datta. "These explanations then inform fixes to the system to avoid future violations."

"The hard part is creating such explanations for systems that employ statistical machine learning," adds Mengshoel. "But doing so is critical, since these methods are increasingly used to power automated decision systems."

But in order to address privacy and fairness in decision systems, the team must first provide formal definitional frameworks of what privacy and fairness truly entail. These definitions must be enforceable and context-dependent, dealing with both protected information itself—like race, gender, or health information—as well as proxies for that information, so that the full scope of risks is covered.

"Committing to philosophical rigor, the project will integrate socially meaningful conceptions of privacy, fairness, and accountability into its scientific efforts," comments Nissenbaum, "thereby ensuring its relevance to fundamental societal challenges."

"Although science cannot decide moral questions, given a standard from ethics, science can shed light on how to enforce it, its consequences, and how it compares to other standards, " says Tschantz.

Another fundamental challenge the team faces is in enabling accountability while simultaneously protecting the system owners' intellectual property, and privacy of the system's users.

"Since accountability mechanisms require some level of access to the system, they can, unless carefully designed, leak the intellectual property of data processors and compromise the confidentiality of the training data subjects, as demonstrated in the prior work of many on the team," says Fredrikson.

"Unfortunately, we don't yet understand what machine learning systems are leaking about privacy-sensitive training data sets. This project will be a great opportunity to investigate the extent to which having access to prediction functions or their parameters reveals sensitive information, and, in turn, how to improve machine learning to be more privacy friendly."

Datta has assembled a truly interdisciplinary team of researchers for the project. Combining the skills of experts in philosophy, ethics, machine learning, security, and privacy, Datta hopes to successfully enable accountability in automated decision systems—an achievement that would add a layer of humanity to artificially intelligent systems.

Provided by Carnegie Mellon University Electrical and Computer Engineering