CRUNCH Seminars at Brown, Division of Applied Mathematics

Friday – January 31, 2020

Paper review: Preventing Undesirable Behavior of Intelligent Machines

Yixiang Deng

Intelligent machines using machine learning algorithms are ubiquitous, ranging from simple data analysis and pattern recognition tools to complex systems that achieve superhuman performance on various tasks. Ensuring that they do not exhibit undesirable behavior—that they do not, for example, cause harm to humans—is, therefore, a pressing problem. We propose a general and flexible framework for designing machine learning algorithms. This framework simplifies the problem of specifying and regulating undesirable behavior. To show the viability of this framework, we used it to create machine learning algorithms that precluded the dangerous behavior caused by standard machine learning algorithms in our experiments. Our framework for designing machine learning algorithms simplifies the safe and responsible application of machine learning.