QHS Special Seminar

Friday, September 8, 2017
3:00PM – 4:00 PM

Albert Sherman Building, AS8-2072

"Safe Machine Learning"

Presented by: Philip Thomas, PhD

Machine learning algorithms are everywhere, ranging from simple data analysis and pattern recognition tools used across the sciences to complex systems that achieve super-human performance on various tasks. Ensuring that they are safe—that they do not, for example, cause harm to humans or act in a racist or sexist way—is therefore not a hypothetical problem to be dealt with in the future, but a pressing one that we can and should address now.

In this talk Dr. Thomas will discuss some of his recent efforts to develop safe machine learning algorithms, and particularly safe reinforcement learning algorithms, which can be responsibly applied to high-risk applications. He will focus on a specific research problem that is central to the design of safe reinforcement learning algorithms: accurately predicting how well a policy would perform if it were to be used, given data collected from the deployment of a different policy. Solutions to this problem provide a way to determine that a newly proposed policy would be dangerous to use without requiring the dangerous policy to ever actually be used.

Philip Thomas is an assistant professor at the University of Massachusetts Amherst. Before that, he was a postdoctoral research fellow in the Computer Science Department at Carnegie Mellon University, advised by Emma Brunskill (2015-2017). He received his Ph.D. from the College of Information and Computer Sciences at the University of Massachusetts Amherst in 2015, where he was advised by Andrew Barto. Prior to that, Philip received his B.S. and M.S. in computer science from Case Western Reserve University in 2008 and 2009, respectively, where Michael Branicky was his adviser. Philip's research interests are in machine learning with emphases on reinforcement learning, safety, and designing algorithms that have practical theoretical guarantees.

Sponsored by: UMMS Department of Quantitative Health Sciences (QHS),
Health Informatics and Implementation Science Division