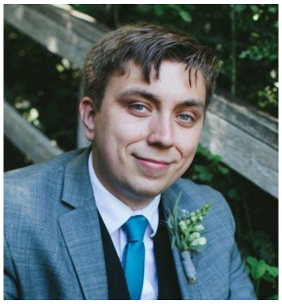
HUMANITY CENTERED ROBOTICS INITIATIVE

BROWN

Daniel Kasenberg
Tufts University

"Toward Agents that Can Learn and Obey Moral Norms"



Thursday, March 14, 2019
12:00-12:50
CIT 477 Lubrano

Abstract: Linear temporal logic (LTL) allows agent designers to specify complex goals and constraints for artificial agents in an interpretable way. This interpretability is especially useful for specifying, e.g., moral and social norms which an agent is expected to obey. In this talk I describe work on algorithms enabling agents in Markov Decision Processes to (1) maximally satisfy a set of LTL objectives which may not all be satisfiable; and (2) infer LTL goals from demonstrated behavior.

Daniel Kasenberg is a Ph.D. candidate in the Human-Robot Interaction Lab at Tufts University. His work is at the intersection of reinforcement learning and machine ethics. His current interest is in developing hybrid agents that can learn, reason about, and obey moral and social norms explicitly specified in logic while leveraging the power of reinforcement learning to act in complex environments.

Host: David Abel/HCRI

For more information on this talk and the HCRI Speaker Series, contact hcri@brown.edu or visit hcri.brown.edu.