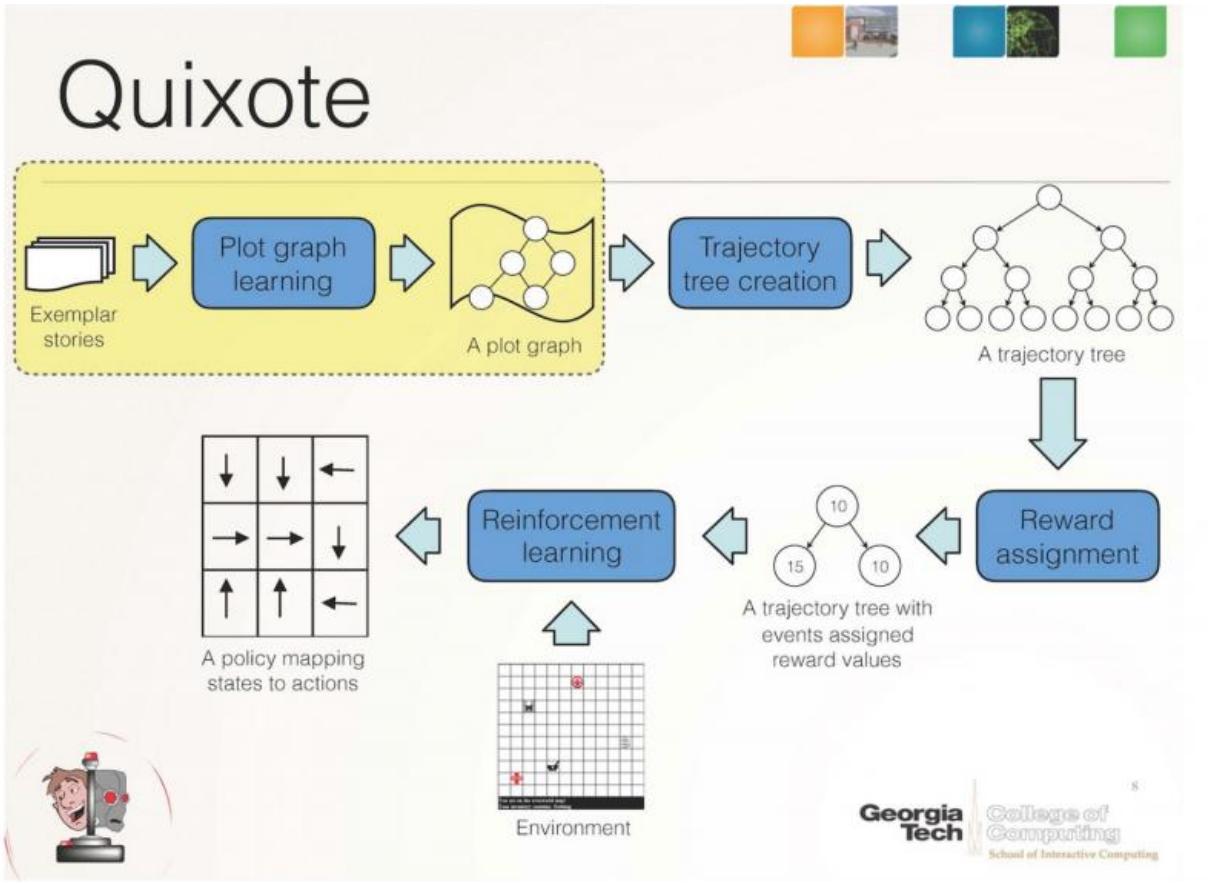
Using stories to teach human values to artificial agents

February 12 2016



The Quixote system by researchers Mark Riedl and Brent Harrison teaches robots how to behave like the protagonist when interacting with humans and is as part of a larger effort to build an ethical value system into new forms of artificial intelligence. Credit: Georgia Tech

The rapid pace of artificial intelligence (AI) has raised fears about whether robots could act unethically or soon choose to harm humans. Some are calling for bans on robotics research; others are calling for more research to understand how AI might be constrained. But how can robots learn ethical behavior if there is no "user manual" for being human?

Researchers Mark Riedl and Brent Harrison from the School of Interactive Computing at the Georgia Institute of Technology believe the answer lies in "Quixote"—to be unveiled at the AAAI-16 Conference in Phoenix, Ariz. (Feb. 12 - 17, 2016). Quixote teaches "value alignment" to robots by training them to read stories, learn acceptable sequences of events and understand successful ways to behave in human societies.

"The collected stories of different cultures teach children how to behave in socially acceptable ways with examples of proper and improper behavior in fables, novels and other literature," says Riedl, associate professor and director of the Entertainment Intelligence Lab. "We believe story comprehension in robots can eliminate psychotic-appearing behavior and reinforce choices that won't harm humans and still achieve the intended purpose."

Quixote is a technique for aligning an AI's goals with human values by placing rewards on socially appropriate behavior. It builds upon Riedl's prior research—the Scheherazade system—which demonstrated how artificial intelligence can gather a correct sequence of actions by crowdsourcing story plots from the Internet.

Scheherazade learns what is a normal or "correct" plot graph. It then passes that data structure along to Quixote, which converts it into a "reward signal" that reinforces certain behaviors and punishes other behaviors during trial-and-error learning. In essence, Quixote learns that it will be rewarded whenever it acts like the protagonist in a story instead

of randomly or like the antagonist.

For example, if a robot is tasked with picking up a prescription for a human as quickly as possible, the robot could a) rob the pharmacy, take the medicine, and run; b) interact politely with the pharmacists, or c) wait in line. Without value alignment and positive reinforcement, the robot would learn that robbing is the fastest and cheapest way to accomplish its task. With value alignment from Quixote, the robot would be rewarded for waiting patiently in line and paying for the prescription.

Riedl and Harrison demonstrate in their research how a value-aligned reward signal can be produced to uncover all possible steps in a given scenario, map them into a plot trajectory tree, which is then used by the robotic agent to make "plot choices" (akin to what humans might remember as a Choose-Your-Own-Adventure novel) and receive rewards or punishments based on its choice.

The Quixote technique is best for robots that have a limited purpose but need to interact with humans to achieve it, and it is a primitive first step toward general moral reasoning in AI, Riedl says.

"We believe that AI has to be enculturated to adopt the values of a particular society, and in doing so, it will strive to avoid unacceptable behavior," he adds. "Giving robots the ability to read and understand our stories may be the most expedient means in the absence of a human user manual."

More information: Paper: www.cc.gatech.edu/~riedl/pubs/aaai-ethics16.pdf

Provided by Georgia Institute of Technology