

DETERMINING ACTION SELECTION POLICIES OF AN EXECUTION DEVICE**ABSTRACT**

Disclosed herein are methods, systems, and apparatus for generating an action selection policy (ASP) of an execution device. One method includes, in a current iteration, computing a first reward for a current state based on respective first rewards for actions in
5 the current state and an ASP of the current state in the current iteration; computing an accumulative respective regret value of each action in the current state based on a difference between the respective first reward for the action and the first reward for the current state; computing an ASP of the current state in the next iteration; computing a second reward for the current state based on the respective first rewards for the actions and the ASP of the
10 current state in the next iteration; and determining an ASP of the previous state in the next iteration based on the second reward for the current state.

FIG. 2