Separate Roles for Reward Magnitude and Uncertainty in the Explore-Exploit Dilemma

Andra Geana, Robert C. Wilson, John Myles White, Elliot A. Ludvig & Jonathan D. Cohen

The world often presents us with multiple potential alternatives. When we drive to visit a friend, we must choose whether to take the familiar route, or try a side road that looks like a shortcut. If we try the side road, we might get there faster, have dinner sooner and get to spend more time with our friend, but we might also get lost and spend our time and gas for nothing. This illustrates the explore-exploit dilemma: the tradeoff between choosing a familiar alternative or searching the environment for other options that may be better or could be worse.

Reward magnitude is a vital factor in balancing exploration and exploitation: organisms allocate more time to, or exploit, the more rewarding alternative. This does not fully account for their behavior, however. They often explore more than reward magnitudes would dictate. More recent work has shown that people may incorporate their level of uncertainty into their decisions, which can bias them toward exploration. An optimal strategy should include this interplay between reward magnitude and uncertainty in balancing exploration and exploitation.

Historically, these two parameters have been confounded. Exploiting the more rewarding option leads to more information about it and lower uncertainty, while the other options are selected less often, leading to higher uncertainty about them. We designed a task that orthogonalizes magnitude and uncertainty and compared how these variables independently affect the explore-exploit tradeoff.

Participants played a two-armed bandit task comprised of 60 games, each containing 15 choices. One of the two bandits provided a certain outcome; the magnitude was displayed on the screen and decreased one point every time it was chosen. The other bandit was uncertain: its payoff was randomly drawn from a Gaussian with a constant mean and variance. By titrating the certain bandit, we estimated people's indifference points (where were equally likely to choose either bandit).

People were sensitive to reward magnitude, exploiting more for higher magnitudes. They were also sensitive to uncertainty, exploring more under high uncertainty. Fitting a softmax psychometric function to the choice data revealed the presence of a significant uncertainty bonus (people chose the uncertain bandit even when its mean was a little worse). This uncertainty bonus is predicted for optimal agents using a normative approach based on dynamic programming. Our data suggests that people may be able to approximate this optimal strategy.