Assessing others: Evaluating the expertise of humans and computer algorithms

December 31 2013, by Cynthia Eller

Two of the images that test subjects saw as they assessed others' expertise: one an image of a person's face, the other an icon said to represent a computer algorithm.

(Medical Xpress)—How do we come to recognize expertise in another person and integrate new information with our prior assessments of that person's ability? The brain mechanisms underlying these sorts of evaluations—which are relevant to how we make decisions ranging from whom to hire, whom to marry, and whom to elect to Congress—are the subject of a new study by a team of neuroscientists at the California Institute of Technology (Caltech).

In the study, published in the journal *Neuron*, Antonio Rangel, Bing Professor of Neuroscience, Behavioral Biology, and Economics, and his associates used functional magnetic resonance imaging (fMRI) to monitor the brain activity of volunteers as they moved through a particular task. Specifically, the subjects were asked to observe the shifting value of a hypothetical financial asset and make predictions about whether it would go up or down. Simultaneously, the subjects interacted with an "expert" who was also making predictions.

Half the time, subjects were shown a photo of a person on their computer screen and told that they were observing that person's predictions. The other half of the time, the subjects were told they were observing predictions from a computer algorithm, and instead of a face, an abstract logo appeared on their screen. However, in every case, the subjects were interacting with a computer algorithm—one programmed to make correct predictions 30, 40, 60, or 70 percent of the time.

Subjects' trust in the expertise of agents, whether "human" or not, was measured by the frequency with which the subjects made bets for the agents' predictions, as well as by the changes in those bets over time as the subjects observed more of the agents' predictions and their consequent accuracy.

This trust, the researchers found, turned out to be strongly linked to the accuracy of the subjects' own predictions of the ups and downs of the asset's value.

"We often speculate on what we would do in a similar situation when we are observing others—what would I do if I were in their shoes?" explains Erie D. Boorman, formerly a postdoctoral fellow at Caltech and now a Sir Henry Wellcome Research Fellow at the Centre for FMRI of the Brain at the University of Oxford, and lead author on the study. "A growing literature suggests that we do this automatically, perhaps even

unconsciously."

Indeed, the researchers found that subjects increasingly sided with both "human" agents and computer algorithms when the agents' predictions matched their own. Yet this effect was stronger for "human" agents than for algorithms.

This asymmetry—between the value placed by the subjects on (presumably) human agents and on computer algorithms—was present both when the agents were right and when they were wrong, but it depended on whether or not the agents' predictions matched the subjects'. When the agents were correct, subjects were more inclined to trust the human than algorithm in the future when their predictions matched the subjects' predictions. When they were wrong, human experts were easily and often "forgiven" for their blunders when the subject made the same error. But this "benefit of the doubt" vote, as Boorman calls it, did not extend to computer algorithms. In fact, when computer algorithms made inaccurate predictions, the subjects appeared to dismiss the value of the algorithm's future predictions, regardless of whether or not the subject agreed with its predictions.

Since the sequence of predictions offered by "human" and algorithm agents was perfectly matched across different test subjects, this finding shows that the mere suggestion that we are observing a human or a computer leads to key differences in how and what we learn about them.

A major motivation for this study was to tease out the difference between two types of learning: what Rangel calls "reward learning" and "attribute learning." "Computationally," says Boorman, "these kinds of learning can be described in a very similar way: We have a prediction, and when we observe an outcome, we can update that prediction."

Reward learning, in which test subjects are given money or other valued

goods in response to their own successful predictions, has been studied extensively. Social learning—specifically about the attributes of others (or so-called attribute learning)—is a newer topic of interest for neuroscientists. In reward learning, the subject learns how much reward they can obtain, whereas in attribute learning, the subject learns about some characteristic of other people.

This self/other distinction shows up in the subjects' brain activity, as measured by fMRI during the task. Reward learning, says Boorman, "has been closely correlated with the firing rate of neurons that release dopamine"—a neurotransmitter involved in reward-motivated behavior—and brain regions to which they project, such as the striatum and ventromedial prefrontal cortex. Boorman and colleagues replicated previous studies in showing that this reward system made and updated predictions about subjects' own financial reward. Yet during attribute learning, another network in the brain—consisting of the medial prefrontal cortex, anterior cingulate gyrus, and temporal parietal junction, which are thought to be a critical part of the mentalizing network that allows us to understand the state of mind of others—also made and updated predictions, but about the expertise of people and algorithms rather than their own profit.

The differences in fMRIs between assessments of human and nonhuman agents were subtler. "The same brain regions were involved in assessing both human and nonhuman agents," says Boorman, "but they were used differently."

"Specifically, two brain regions in the prefrontal cortex—the lateral orbitofrontal cortex and medial prefrontal cortex—were used to update subjects' beliefs about the expertise of both humans and algorithms," Boorman explains. "These regions show what we call a 'belief update signal.'" This update signal was stronger when subjects agreed with the "human" agents than with the algorithm agents and they were correct. It

was also stronger when they disagreed with the computer algorithms than when they disagreed with the "human" agents and they were incorrect. This finding shows that these brain regions are active when assigning credit or blame to others.

"The kind of learning strategies people use to judge others based on their performance has important implications when it comes to electing leaders, assessing students, choosing role models, judging defendents, and so on," Boorman notes. Knowing how this process happens in the brain, says Rangel, "may help us understand to what extent individual differences in our ability to assess the competency of others can be traced back to the functioning of specific brain regions."

Provided by California Institute of Technology