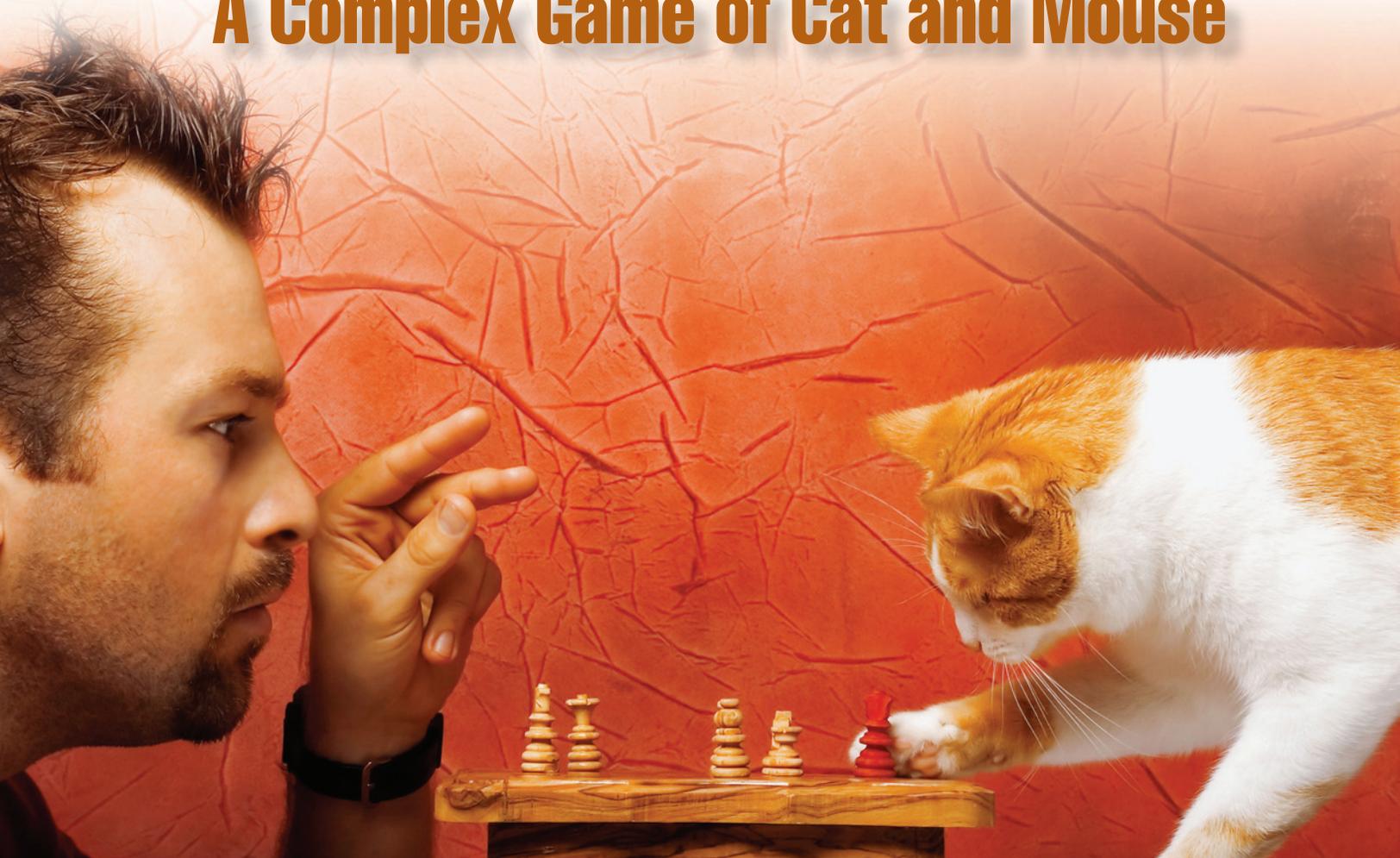


# A Complex Game of Cat and Mouse



**I**n chess, whether one is playing against a human or a computer, each move typically builds on the opponent's previous maneuvers. Humans have one significant advantage, however: their innate ability to analyze many factors before taking a turn. For example, human players may consider what they know about an opponent's behaviors or past strategies. They then use this knowledge to devise a game plan that gives them the advantage.

Traditional artificial intelligence systems, even those that model real-world adversarial scenarios, are less adaptive. In these systems, players, or "agents," are programmed to process a far less nuanced, narrow set of conditions before making a decision, and they often select moves with little or no regard for the behavior of others in the game environment.

As part of a two-year project funded by the Laboratory Directed Research and Development Program, a research team in Livermore's Engineering Directorate is attempting to develop

computer modeling systems that process decisions more like humans do, adapting to the changing environment. The team, which includes computer scientists Brenda Ng and Kofi Boakye, operations researcher Carol Meyers, and former summer intern Andrew Wang from the Massachusetts Institute of Technology, wants to devise a system that can analyze adversarial relationships for a wide range of national security and law-enforcement applications. "The idea is to create a framework that takes into account an agent's intent rather than simply its static behavior," says Ng, who leads the project. This type of system will improve the effectiveness of computer simulations designed to analyze response scenarios against real-world adversaries.

## Outwitting the Enemy

In single-agent decision models, the agent can choose from a given set of actions to advance from one state to another within the

simulation environment. The agent selects an action based on the rewards or penalties it will incur for that turn. Most of the time, the information available for evaluating these options is imperfect. That is, the agent cannot be certain about its current state before deciding which optimal action to take. In these situations, the agent must infer its state through the noisy observations it receives from the environment. To do this, the agent maintains beliefs, in the form of probability distributions, over all of its possible states.

“This process is analogous to attempting to optimize a retirement portfolio,” says Ng. “One can only surmise through day-to-day reports (observations) the general health of his or her investment (the agent state). Possible actions are to buy, sell, or cash out, but each action has an associated cost—a reward or a penalty.”

As the number of agents increases, the models become more complex. The interactive, partially observable Markov decision process (I-POMDP) model is well suited for adversarial scenarios between multiple agents because it allows agents to consider the capabilities and beliefs of their adversaries before making the next move. Within such an environment, agents repeatedly interact with one another, and each agent’s actions affect the joint state of all agents, which in turn affects every agent’s observations.

One drawback with the I-POMDP model is with the built-in assumption that agents know all of the model parameters. In the real world, many conditions remain unknown until people interact with each other, whether they are allies or adversaries. To make the agents’ simulated behavior more realistic, the Livermore team incorporated reinforcement learning into the I-POMDP model.

With the new framework, agents learn as they make choices within the established environment. Each interaction provides information that helps them select the optimal action for a given situation, allowing the agents to maximize their rewards. Model parameters are not fully known beforehand, but agents learn them through trial and error as the players interact.

“Our goal is to bridge the gap between theory and practice in what an I-POMDP can model in an adversarial scenario,” says Meyers. A framework that simulates how agents “learn” from their opponents and change strategies based on observed behavior has major potential for law-enforcement and national security applications.

### Show Me the Money

During the project’s first year, the team applied the conventional I-POMDP to a simplified money-laundering scenario to evaluate its potential for accurately modeling dynamic adversaries. Considered a high-stakes game of cat and mouse, money laundering typically involves a complex series of financial transactions intentionally designed to be difficult to trace. Adversaries who think their actions are being monitored are likely to change their behavior, taking steps to evade detection or deceive the other agent.



Interactive, partially observable Markov decision process (I-POMDP) models incorporate the idea of nested belief to emulate adversarial relationships between multiple agents. In I-POMDP models, each agent tries to anticipate an opponent’s behavior and makes choices to counter those actions.

“The money-laundering scenario is appealing because both agents have nested beliefs,” says Ng. “Each one acts on what it ‘believes’ the other is thinking.” The nested-belief framework attempts to model each player’s thought processes and actions in a manner that better simulates human behavior—what Ng calls an I-think-that-you-think-that-I-think pattern.

The team’s initial model consisted of two agents. The first agent, a money launderer, is trying to diffuse and integrate its assets, “dirty” money, into the mainstream economy without being detected. The second agent, a law-enforcement officer, wants to confiscate this money before the money launderer can “cash out” via transactions with legitimate businesses.

The two agents operate within a defined number of states where the laundered money may be placed or found. For the money launderer, each state represents a location through which money can be diverted, such as bank accounts, trusts, or securities, as well as businesses that can integrate the large sums, for example,

casinos and real-estate agencies. For law enforcement, each state represents a location where the officer can probe for suspicious activities. Both agents take actions not only to gather intelligence information on the opponent but also to transition from state to state. The “game” resets when the money is either successfully laundered or confiscated.

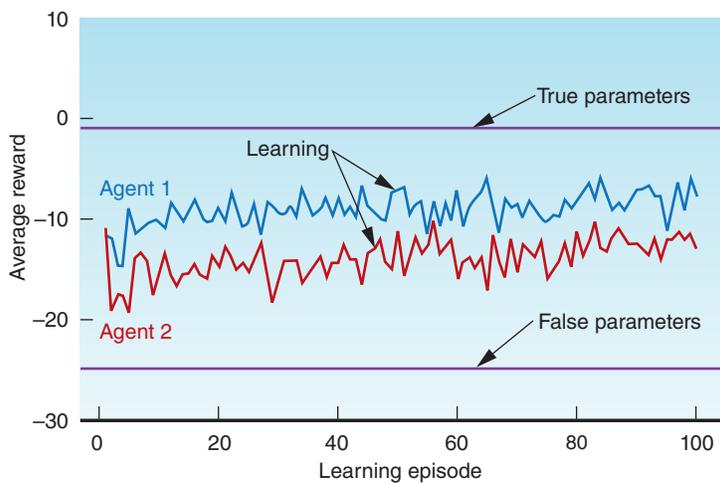
One challenge for the Livermore team was reducing the computational burden required to calculate the solution. “We had to substantially scale down our model to make it run efficiently,” says Meyers. “Even in the simplified version, the number of actions, observations, and states was 20 times greater than that in any game previously solved by an I-POMDP.” The researchers also modified algorithms designed to solve the I-POMDP models. They introduced additional approximations to a technique

called reachability tree sampling, in which possible paths forward in the game are based on the likelihood of each path. These modifications are designed to focus the agents’ attention only on the highly probable observations as determined by computing the optimal actions.

Throughout an I-POMDP simulation, both agents maintain beliefs about the physical states of their environment and are knowledgeable of model parameters, such as how the other’s actions will contribute to the next state. After experimenting with the model, the Livermore team determined that under most conditions, the money launderer has the advantage. However, when both agents are set to focus on achieving immediate rewards, the law-enforcement officer wins more often and does so much faster.

The tiger problem is a standard benchmark used to evaluate agent modeling and decision-making frameworks. In the two-player version of this game, agents in identical scenarios on adjacent floors must decide which door to open. Behind one door is a jackpot, but behind the other is a tiger. The agents can take one of three actions: open the left door, open the right door, or listen in an attempt to learn from the other agent’s choice. (Rendering by Sabrina Fletcher.)





This graph compares the rewards acquired by two agents using the Livermore Bayes-adaptive I-POMDP model, which allows agents to learn the model parameters, and the conventional I-POMDP model, which bases agent decisions on known parameters that are either true or false.

### A Risky Proposition

In the second phase of the project, the team adapted the lessons learned from the money-laundering experiment to create a Bayes-adaptive I-POMDP framework. Standard I-POMDPs can model only the agents' attempts to reason. Bayes-adaptive I-POMDPs, however, can also model their attempts to learn. In this new framework, agents interact with their opponents to acquire information about their state and the dynamics governing states and observations. The Bayes-adaptive I-POMDP thus improves results produced when modeling human adversarial relationships.

"We assume that the state, action, and observation spaces are finite and known, but the model parameters—namely, the probabilities with which the agents change states and get specific information—are not fully known," says Ng. "This approach is more realistic because in the real world, agents would face a number of uncertainties as both parties try to deceive each other."

To demonstrate how adversaries continually adapt to an opponent, the team applied the model to the tiger problem, a standard benchmark used in academia. In the two-agent scenario, two adjacent rooms contain an object, either a ferocious tiger or a jackpot. The two agents have access to their own set of doors and can hear but not see the other agent. Each agent can take one of three actions: open the left door, open the right door, or listen.

An agent choosing to listen might hear a tiger growl, a door creak, or only silence. However, observations are obscured by background noise, so the agent cannot completely trust what it hears. After listening, the agent can update its belief state, learn about the truthfulness of the observations, and then choose the next optimal action. At the same time, both agents are trying to anticipate the action, observation, and evolving belief of the opponent.

The model now has more states to track because the state space includes parameters that enable learning. As a result, the team had to add further algorithmic approximations. "We transferred our approximations from the money-laundering model," says Boakye, "and then revised them to work for the larger state space." The tiger simulations revealed that when both agents are learning, the agents reap more rewards as the accuracy of their learned parameters increases. In essence, the learned behavior allows the agent to significantly improve its rewards compared with those attained from an incorrect model with no learning. In addition, when both agents are learning, rewards take longer to acquire, which is similar to a real scenario in which adversaries try to "game" each other.

### I See You

Ng cautions that although the team's demonstration was successful, the framework does not yet provide a complete platform for modeling complicated human adversarial systems. "Even more algorithmic approximations would be required," she says. "Our work does however provide a major advance in fundamental adversarial modeling. It has great promise for a variety of national security applications, including counterterrorism efforts."

Ongoing research will focus on developing ways to enable more states, actions, and observations in the model while keeping the computation tractable. With more realistic adversarial models in the works, national security and law-enforcement officials may one day have a better tool for understanding their intelligent opponents. As a result, these systems may also help answer a fundamental question: how might adversaries act differently if they knew they were being watched?

—Caryn Meissner

**Key Words:** adversarial modeling; artificial intelligence; counterterrorism; interactive, partially observable Markov decision process (I-POMDP); law enforcement; money laundering; reinforcement learning.

**For further information contact Brenda Ng (925) 422-4553 (ng30@llnl.gov).**

