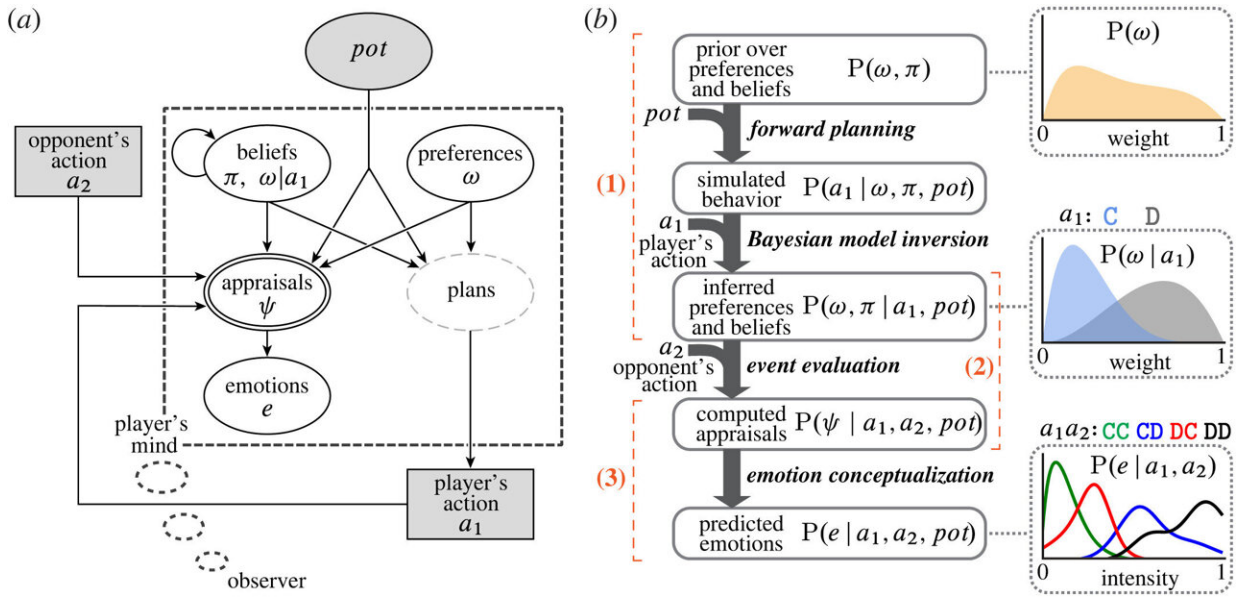
Computational model mimics humans' ability to predict emotions

June 5 2023, by Anne Trafton



Emotion prediction as inference over an intuitive Theory of Mind. Hypotheses about how human observers reason about others' emotions can be formalized as probabilistic generative models. This reflects a hypothesis about observers' intuitive theory of other people's minds, not a scientific hypothesis about people's actual emotions. (*a*) Implementation of the general hypothesis for the 'Split or Steal' game (a public one-shot Prisoner's Dilemma). We treat observers' emotion predictions as a function of their intuitive reasoning about how players will subjectively evaluate, or 'appraise', the game's outcome. Observers predict a player's emotions by inferring what preferences and beliefs motivated the player's decision to Cooperate or Defect, and reason about how those preferences and beliefs would cause the player to emotionally react to the outcome of the game. The intuitive theories we test take the form of directed acyclic graphs, where arrows indicate the causal relationship between variables.

Shaded nodes are observable variables and open nodes are latent variables. Round nodes are continuous variables, rectangular nodes are discrete variables. Nodes with a single border are random variables. The double border indicates that appraisals are calculated deterministically. Plans are shown with a partial border because they are not explicitly represented in this model. (*b*) Computational model of the intuitive theory. The model comprises three modules. Module (1) infers a joint distribution over preferences and beliefs given a player's action via inverse planning. Module (2) computes appraisals based on how a player would evaluate the outcome of the game with respect to the inferred preferences and beliefs. Module (3) generates emotion predictions by transforming the computed appraisals. The probability density plots illustrate how observers' prior belief about a player's preference P(ω) is updated based on the player's action, and how the inferred preference P(ω|a1) is used to predict the player's emotional reaction to the game's outcome P(e|a1,a2). Credit: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (2023). DOI: 10.1098/rsta.2022.0047

When interacting with another person, you likely spend part of your time trying to anticipate how they will feel about what you're saying or doing. This task requires a cognitive skill called theory of mind, which helps us to infer other people's beliefs, desires, intentions, and emotions.

MIT neuroscientists have now designed a computational model that can predict other people's emotions—including joy, gratitude, confusion, regret, and embarrassment—approximating human observers' social intelligence. The model was designed to predict the emotions of people involved in a situation based on the prisoner's dilemma, a classic game theory scenario in which two people must decide whether to cooperate with their partner or betray them.

To build the model, the researchers incorporated several factors that have been hypothesized to influence people's emotional reactions,

including that person's desires, their expectations in a particular situation, and whether anyone was watching their actions.

"These are very common, basic intuitions, and what we said is, we can take that very basic grammar and make a model that will learn to predict emotions from those features," says Rebecca Saxe, the John W. Jarve Professor of Brain and Cognitive Sciences, a member of MIT's McGovern Institute for Brain Research, and the senior author of the study.

Sean Dae Houlihan Ph.D. '22, a postdoc at the Neukom Institute for Computational Science at Dartmouth College, is the lead author of the paper, which appears today in *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. Other authors include Max Kleiman-Weiner Ph.D. '18, a postdoc at MIT and Harvard University; Luke Hewitt Ph.D. '22, a visiting scholar at Stanford University; and Joshua Tenenbaum, a professor of computational cognitive science at MIT and a member of the Center for Brains, Minds, and Machines and MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL).

Predicting emotions

While a great deal of research has gone into training computer models to infer someone's emotional state based on their facial expression, that is not the most important aspect of human emotional intelligence, Saxe says. Much more important is the ability to predict someone's emotional response to events before they occur.

"The most important thing about what it is to understand other people's emotions is to anticipate what other people will feel before the thing has happened," she says. "If all of our emotional intelligence was reactive, that would be a catastrophe."

To try to model how human observers make these predictions, the researchers used scenarios taken from a British game show called "Golden Balls." On the show, contestants are paired up with a pot of $100,000 at stake. After negotiating with their partner, each contestant decides, secretly, whether to split the pool or try to steal it. If both decide to split, they each receive $50,000. If one splits and one steals, the stealer gets the entire pot. If both try to steal, no one gets anything.

Depending on the outcome, contestants may experience a range of emotions—joy and relief if both contestants split, surprise and fury if one's opponent steals the pot, and perhaps guilt mingled with excitement if one successfully steals.

To create a computational model that can predict these emotions, the researchers designed three separate modules. The first module is trained to infer a person's preferences and beliefs based on their action, through a process called inverse planning.

"This is an idea that says if you see just a little bit of somebody's behavior, you can probabilistically infer things about what they wanted and expected in that situation," Saxe says.

Using this approach, the first module can predict contestants' motivations based on their actions in the game. For example, if someone decides to split in an attempt to share the pot, it can be inferred that they also expected the other person to split. If someone decides to steal, they may have expected the other person to steal, and didn't want to be cheated. Or, they may have expected the other person to split and decided to try to take advantage of them.

The model can also integrate knowledge about specific players, such as the contestant's occupation, to help it infer the players' most likely motivation.

The second module compares the outcome of the game with what each player wanted and expected to happen. Then, a third module predicts what emotions the contestants may be feeling, based on the outcome and what was known about their expectations. This third module was trained to predict emotions based on predictions from human observers about how contestants would feel after a particular outcome.

The authors emphasize that this is a model of human social intelligence, designed to mimic how observers causally reason about each other's emotions, not a model of how people actually feel.

"From the data, the model learns that what it means, for example, to feel a lot of joy in this situation, is to get what you wanted, to do it by being fair, and to do it without taking advantage," Saxe says.

Core intuitions

Once the three modules were up and running, the researchers used them on a new dataset from the game show to determine how the models' emotion predictions compared with the predictions made by human observers. This model performed much better at that task than any previous model of emotion prediction.

The model's success stems from its incorporation of key factors that the human brain also uses when predicting how someone else will react to a given situation, Saxe says. Those include computations of how a person will evaluate and emotionally react to a situation, based on their desires and expectations, which relate to not only material gain but also how they are viewed by others.

"Our model has those core intuitions, that the mental states underlying emotion are about what you wanted, what you expected, what happened, and who saw. And what people want is not just stuff. They don't just

want money; they want to be fair, but also not to be the sucker, not to be cheated," she says.

"The researchers have helped build a deeper understanding of how emotions contribute to determining our actions; and then, by flipping their model around, they explain how we can use people's actions to infer their underlying emotions. This line of work helps us see emotions not just as 'feelings' but as playing a crucial, and subtle, role in human social behavior," says Nick Chater, a professor of behavioral science at the University of Warwick, who was not involved in the study.

In future work, the researchers hope to adapt the model so that it can perform more general predictions based on situations other than the game-show scenario used in this study. They are also working on creating models that can predict what happened in the game based solely on the expression on the faces of the contestants after the results were announced.

More information: Sean Dae Houlihan et al, Emotion prediction as computation over a generative theory of mind, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* (2023). DOI: 10.1098/rsta.2022.0047