Chapter 14

Theory of mind and inverse decision-making

Julian Jara-Ettinger, Chris Baker, Tomer Ullman, & Joshua B. Tenenbaum

To effectively interact with other people, we must continually infer and monitor their mental states: what they think, what they want, and what they know about the world – and what they think, want and know about our own mental states and those of others. Even as passive observers, the ability to understand other people's behavior in terms of mental states provides a powerful tool for social learning. Watching a more knowledgeable person's behavior can reveal to us how the world works. Watching a more experienced person can teach us when persistence or practice is helpful and when it's not. Attending to how people act towards others can let us determine who is nice, opportunistic, or mean, and how we ourselves should act to be (and be perceived as) positive social partners.

Whenever we explain, predict, or judge each other's behavior, we do so by thinking about their minds. Yet, other people's minds are unobservable, making the ability to infer mental states from observable actions a pre-requisite for human-like social intelligence. This capacity is known as a **Theory of Mind**. The hypothesis of this chapter is that these abilities in humans can be understood as approximate Bayesian inferences over a mental model of how people think and act. We show how the same Bayesian framework used to model rational action planning (Chapter 7) can also be used to model other people's mental processes and infer their latent mental states, as a form of **inverse planning**. In contrast to Chapter 7, which focuses on generating high-value actions given a world model and a utility function, here we focus on attributing a world model and utility function to other agents, with the goal of explaining their behavior under the assumption that they are planning rationally: choosing actions that they expect to have high value given their world models and utility functions which we seek to infer.

14.1 Representing and inferring desires

We begin by considering one of the simplest social situations: watching someone with perfect knowledge choose an option from a finite set of possibilities. (We'll consider more complex cases below, where knowledge is less than perfect, or choices unfold over a sequence of actions, and a range of other real-world complications come into play.) Imagine, for instance, watching a friend choose between chocolate cake and ice cream for dessert. Intuitively, your friend's choice (an observable action) reveals their preference (a mental state). In this setting, inverse planning reduces to a simpler form of inverse decision-making.

To formalize this intuition, we can begin by defining an event as a set of possible world states and a set of possible actions that an agent can take to change the state of the world. In the example above, the state space is $\mathcal{S} = \{\emptyset, \text{Cake}, \text{Ice cream}\}$, which consists of the state where your friend doesn't have dessert ($\emptyset$), the state where your friend has cake (Cake), and the state where your friend has ice cream (Ice cream). The action space is $\mathcal{A} = \{$"order cake," "order ice cream"$\}$. Your friend's preferences can then be represented by a reward function $R : \mathcal{S} \to \mathbb{R}$ that associates each state of the world with a scalar which can be positive (meaning the agent *likes* the state) or negative (meaning the agent *dislikes* the state). In this situation, we can make four assumptions. First, you and your friend know the current state of the world, which is initially $\emptyset$, as your friend hasn't ordered dessert yet. Second, you and your friend also know the state space and the action space (i.e., the dessert options and the fact that these can be ordered is known). Third, actions are always successful: Taking the actions $a =$ "order cake" and $a =$ "order ice cream", always result in having cake ($s =$"cake") and ice cream ($s =$"ice cream"), respectively. Finally, we'll assume that only your friend know their own rewards (i.e., how much they like ice cream and cake).

After your friend takes observable action $a \in A$, we can infer their underlying reward function by computing the posterior probability through Bayesian inference:

$$p(R|a) \propto p(a|R)p(R). \tag{14.1}$$

Here, $p(R)$ is the observer's prior distribution over possible reward functions, capturing the observer's expectations about what other people generally like. $p(a|R)$ is the probability that your friend would

Figure 14.1: Posterior distribution over the reward of ice cream and cake after watching someone choose ice cream. Probabilities are represented with a graded color range from dark blue (lower probabilities) to bright yelow (high probabilities). a) Posterior distribution using the deterministic likelihood function (Equation 14.3). b) Posterior distribution using the softmaxed likelihood function (Equation 14.4 with $\beta = 0.25$).

take action a if their preferences were correctly represented by reward function R. If we assume that our friend's attitude towards cake is independent of their attitude towards ice cream, we can treat each component of the reward function as independent. That is, the prior probability over any combination of preferences for cake and ice cream is given by the prior probability over ice cream rewards times the prior probability over cake rewards. Formally:

$$p(R) = \prod_{s \in S} p(R(s)) \tag{14.2}$$

To compute the probability $p(a|R)$ of an observed action given a reward function we need a model of how people act. Empirical data with children and adults suggests that, in a situation like this one, people expect agents to take the action leading to the highest possible reward (Lucas et al., 2014; Jern, Lucas, & Kemp, 2017). This can be captured through a simple decision model where

$$a = \begin{cases} \text{order cake}, & \text{if } R(\text{cake}) > R(\text{ice cream}) \\ \text{order ice cream}, & \text{if } R(\text{cake}) < R(\text{ice cream}) \\ \text{Bernoulli}(0.5), & \text{otherwise} \end{cases} \tag{14.3}$$

Under this model, $p(a|R)$ is 1 whenever a selects the state with the highest reward, 0 when it does not, and 0.5 when the two rewards are identical. Figure 14.1a shows the posterior distribution over your friend's preferences after they take action $a =$ "order ice cream" using this simple decision model, and a uniform prior on rewards over the range $[0, 1]$.

In practice, models that expect agents to always maximize rewards are unrealistically strict. For instance, suppose that your friend likes both ice cream and cake, but has a very small preference for

ice cream. According to a strict reward-maximizing model, your friend should order ice cream every single time they face this choice, but intuitively, we would expect your friend to sometimes order cake instead. This can be accounted for by relaxing the expectation that agents strictly maximize rewards to an expectation that agents probabilistically maximize rewards. We can achieve this by building a likelihood function that applies a softmax function to the agent's rewards:

$$p(a|R) \propto \exp(\beta R(s_a)). \tag{14.4}$$

Here, a is the agent's observable action, R is the unobservable reward function guiding their action, and s_a is the resulting state (e.g., if $a =$"order cake," then $s_a =$"cake"). This element of randomization is also known as adopting the **Boltzmann policy** for selecting actions (Sutton & Barto, 1998).

As Equation 14.4 shows, the softmax is a simple transformation where each value (in this case, the rewards $R(s_a)$) is multiplied by a scalar (the parameter β) and exponentiated. After applying this transformation to all values, the full list is normalized (by dividing the right term by a constant $\sum_{a' \in A} \exp(\beta R(s_{a'}))$, so that the terms for choosing the different action sum to 1). This process transforms a list of scalars (in this case, rewards) into a probability distribution over actions. The "rationality" behind this transformation is modulated by the temperature parameter $\beta \in [0, \infty)$. The higher β is, the more the resulting distribution concentrates probability on the options with the highest possible rewards. As $\beta \to \infty$, the probability in $p(a|R)$ increasingly concentrates on the action associated with the highest possible reward, converging to the deterministic reward-maximizing model from Equation 14.3. In contrast, the lower that β is, the resulting distribution spreads the probability across all of the options, while still assigning a higher probability to options with higher values. At the limit, when $\beta = 0$, $p(a|R)$ becomes a uniform distribution over actions, expressing the idea that agents do not act in response to their rewards at all. Thus, β allows us to relax the expectation that agents strictly maximize rewards to an expectation that agents probabilistically maximize rewards (by decreasing the value of β).

Figure 14.1b shows the posterior distribution over reward functions that take values in the range $[0, 1]$ after we watch our friend choose ice cream, using $\beta = 0.25$. As this figure shows, inference over this probabilistic model shows a graded inference. Our initial model (Figure 14.1a) judged that any reward where R(ice cream)$>$ R(cake) was equally probable. By contrast, our softmaxed model (Figure 14.1b) now believes that R(ice cream) is much higher than R(cake) are more likely. This is because, intuitively, a weaker preference for ice cream would give the agent a higher chance of choosing to eat cake (which we did not observe).

Jern et al. (2017) showed how this approach produces human-like preference inferences. In one of their tasks, participants watched an agent choose between different meals, each consisting of multiple food items. In Figure 14.2a, for instance, the agent could take an eggplant dish and a cookie, a chicken dish and a slice of cake, or a fish dish and an apple. After watching the agent choose the eggplant dish with the cookie, participants were asked to infer the agent's preference for different food items. As Figure 14.2b shows, people's inferences correlated highly with the inferred rewards using the Bayesian framework (where the reward associated with each food option was given by the sum of the rewards of each food item; see the schematic of the full space of stimuli in Figure 14.2c). Further, this same model unifies a range of inferences that young children make (Lucas et al., 2014) suggesting that these inferences are at work from early in childhood.

People's actions usually incur a cost (in terms of time and physical effort), and people's mental-state inferences account for how these costs affect agents' choices (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). We can capture this by extending our framework to include cost functions and utility functions. A cost function $C : \mathcal{A} \to \mathbb{R}^+$ is a mapping from actions to positive scalar values that represent negative consequences associated with taking different actions. As we will see in Section 14.4, this term can capture highly abstract aspects of cost, but we can begin by thinking of costs in terms of money (in economic contexts) or energy (in biological contexts).

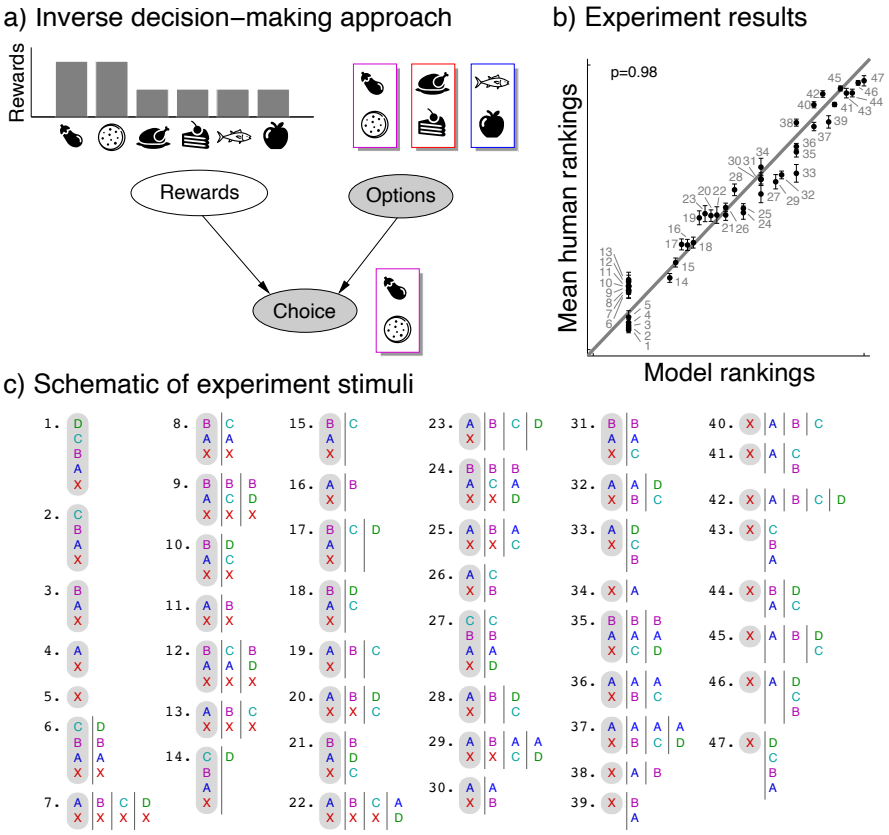

Figure 14.2: Reward inferences from Jern et al. (2017; Positive attributes condition of Experiment 1). a) Schematic of task and model setup. The space of available choices, and the choice that is selected, are observable, and the participant's task is to infer the reward associated with different features (in this case, the different food items) of the choice. b) Experiment results. Each point represents a trial with the model prediction on the horizontal axis and participant judgments on the vertical axis. c) Experimental stimuli. Trials are numbered 1-47. In each trial, each column represents a potential option. Within each choice, letters represent different features. The shaded columns show the agent's selected choice. Reproduced with permission from Jern et al. (2017).

A utility function $U : \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ is a mapping that associates every possible combination of states and actions with the difference between the incurred costs and attained rewards,

$$U(a, s) = R(s) - C(a). \tag{14.5}$$

This utility function captures the expectation that agents value action plans that yield high rewards while incurring the lowest possible costs (Jara-Ettinger et al., 2016; Liu, Ullman, Tenenbaum, & Spelke, 2017; Csibra, Bíró, Koós, & Gergely, 2003).

If the action costs are known, we can infer agents' unobservable rewards by now assuming that agents' act to probabilistically maximize their utilities (rather than just their rewards), using the likelihood function

$$p(a|R; C) \propto \exp(\beta U(a, s_a)) \tag{14.6}$$

where the left hand side of the equation is the probability of choosing an action, given a specific reward

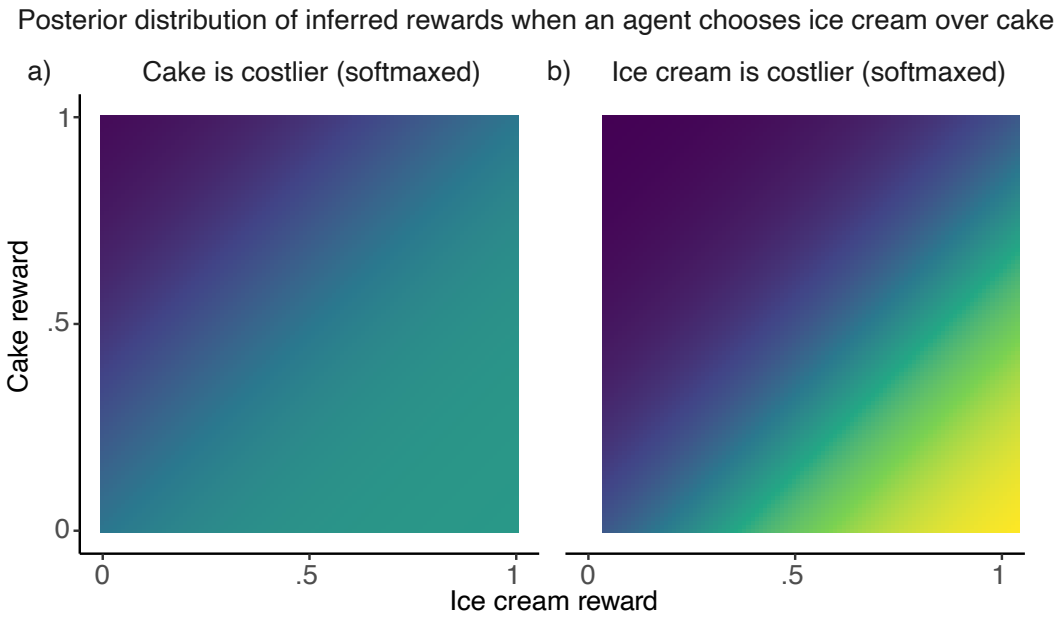

Figure 14.3: Posterior distribution over the reward of ice cream and cake after watching someone choose ice cream, in a context where costs are at play. a) The posterior distribution using a softmaxed likelihood function ($\beta = 0.25$) when cake incurs a cost of 0.25 and ice cream has cost 0. b) The posterior using a softmaxed likelihood function ($\beta = 0.25$) when ice cream incurs a cost of 0.25 and cake has cost of 0.

and cost function, and right side of the equation is the softmax of the utility function (which is in turn just the reward minus the cost).

Figures 14.3a-b show the posterior distribution over reward functions when your friend chooses ice cream in a context where each choice incurs a different cost. When cake is more costly, seeing your friend choose ice cream no longer implies that the reward for ice cream is higher than the reward for cake (Figure 14.3a); your friend might have preferred it because the cost was lower (therefore making the utility higher). Conversely, when ice cream is more costly, seeing your friend order it provides even stronger evidence that they prefer ice cream over cake. After all, the reward must have been high enough to justify the additional cost. This inference is shown in Figure 14.3b, where the posterior distribution now concentrates the probability on regions where the reward for ice cream is much higher than the reward for cake. As we've not, cost need not be money—represents any factors that an agent might find aversive.

In many situations, however, we do not know other people's costs. In these cases, it is possible to jointly infer an agent's costs and rewards from their choices, via Bayesian inference

$$p(R, C|a) \propto p(a|R, C)p(R, C), \tag{14.7}$$

where the priors over cost and reward can be assumed to be independent, such that $p(R, C) = p(R)p(C)$, and the likelihood function is given by Equation 14.6. In principle, any sequence of actions can be explained by many (infinite, in fact) different combinations of costs and rewards. We revisit this in Section 14.3, where we show how spatial information, priors over costs and rewards, and access to multiple observations constrain these inferences and make the problem tractable.

14.2 Representing and inferring beliefs

Many situations involve reasoning about agents acting under incomplete or incorrect knowledge, and interpreting their behavior involves inferring what they know or believe.

14.2.1 Beliefs about costs and rewards

Continuing with the dessert-selection example (whether to order cake or ice cream), suppose that your friend isn't sure about how much they will like each option. In this case, we cannot represent an agent as having a single reward associated with each dessert. Instead, we can capture their uncertainty about their rewards using probability distributions.

To illustrate how, we'll begin by assuming that the range of possible rewards is finite, falling in the $[0, 1]$ range (although the framework can trivially be extended to infinite reward ranges). In this range of rewards, a wide range of possible beliefs can be represented through Beta distributions (see Chapter 3; the logic presented here can be applied to any parameterized probability distribution). Because the shape of Beta distributions are entirely determined by two parameters—α and β—inferring your friend's beliefs becomes equivalent to inferring these two parameters. Formally, if $b_{ic} = \{\alpha_{ic}, \beta_{ic}\}$ and $b_c = \{\alpha_c, \beta_c\}$ represent your friend's beliefs about how much they will enjoy ice cream (b_{ic}) and cake (b_c), respectively, the posterior over their beliefs about their rewards is given by Bayes' rule:

$$p(b_{ic}, b_c|a) \propto p(a|b_{ic}, b_c)p(b_{ic}, b_c). \tag{14.8}$$

To set the prior distribution $p(b_{ic}, b_c)$, we can assume that your friend's belief about how much they'll enjoy ice cream is independent of their belief about how much they'll enjoy cake, so that $p(b_{ic}, b_c) = p(b_{ic})p(b_c)$. Note that because b_{ic} and b_c represent probability distributions (although each distribution technically consists of two parameters, α and β), their priors consist of a mapping that assigns a probability to each possible probability distribution. That is, these priors capture our belief that your friend might have different kinds of beliefs about their rewards (e.g., we might assign a low prior probability to distributions that reflect a belief that the desserts have a low reward, and a higher prior probability to distributions that reflect a belief that the desserts have a high reward). Each of these prior distributions can be set by assigning a prior distribution to the parameters α and β. Because these two parameters can take any value in range $(0, \infty)$, the prior can be represented through any probability distribution defined over positive real-numbers, such as an exponential or a gamma distribution (see Chapter 3).

The likelihood can then be computed by integrating over the possible rewards your friend expects to obtain:

$$p(a|b_{ic}, b_c) = \int_{R_{ic}=0}^{1} \int_{R_c=0}^{1} p(a|R_{ic}, R_c)p(R_{ic}, R_c|b_{ic}, b_c)\, dR_{ic}\, dR_c, \tag{14.9}$$

where $p(a|R_{ic}, R_c)$ is the probability that your friend would take observed action a if the rewards for ice cream and cake were R_{ic} and R_c (computed using a softmax choice model from Equation 14.4), and $p(R_{ic}, R_c|b_{ic}, b_c)$, is your friend's belief that each dessert will yield these rewards:

$$p(R_{ic}, R_c|b_{ic}, b_c) = \text{Beta}(R_{ic}; b_{ic})\text{Beta}(R_c; b_c). \tag{14.10}$$

In the generalized case with m options, computing the posterior distribution becomes prohibitively expensive but can be solved through sampling-based methods (Chapter 6). This framework can be trivially extended to include inferences about beliefs about costs (see Jara-Ettinger, Floyd, Tenenbaum, & Schulz, 2017, for an applied model that shows how young children's mental-state inferences about others can be explained by a model that accounts for the possibility that agents can be uncertain about their own costs and rewards).

14.2.2 Uncertainty over states of the world and world dynamics

So far, we have assumed that each action is deterministically associated with a corresponding state (e.g. action 'Order cake' leads to the state 'cake'). This is rarely the case, as people's actions can fail to have their intended consequence. If people consider the chance that their actions will be successful when deciding what to do, then our inferences about their behavior must account for this. To achieve this, we can introduce an uncertainty model through a transition function $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$, where $T(s, a, s')$ is the agent's belief that taking action a in state s will change the state to s'. Notice that this transition function can either express the true probabilistic structure of the environment, or simply the agent's uncertainty about the structure of the environment.

Returning to our dessert example, suppose that your friend notices that the waiter is extremely busy and might forget their order. Because the cake takes more time to prepare, ordering it increases the chance that the waiter will forget about it. If the waiter has a 20% chance of forgetting to bring ice cream, and a 40% chance of forgetting to bring cake, we can represent your friend's expectations through the following transition function T:

$a =$ ice cream

	$\emptyset$	cake	ice cream
$\emptyset$	0.2	0	0.8
cake	0	0.2	0.8
ice cream	0	0	1

$a =$ cake

	$\emptyset$	cake	ice cream
$\emptyset$	0.4	0.6	0
cake	0	1	0
ice cream	0	0.6	0.4

In each matrix, entry (i, j) shows the probability of switching from state i to state j when taking the action "order ice cream" (left matrix) and "order cake" (right matrix). For instance, the first row of the first matrix indicates that if your friend doesn't have dessert yet ($s = \emptyset$) and orders ice cream, there is a 20% chance that they will not get any dessert, a 0% chance that they will get cake, and an 80% chance that they will get ice cream. Similarly, the second row indicates that your friend has cake and orders ice cream, there is a 0% chance that they will be left without dessert (as they already have cake), a 20% chance that the waiter will forget and they will be left with cake, and an 80% chance that the waiter will replace the cake with ice cream.

Under this world model, your friend might order cake because they do not want to risk being left without any dessert. To formalize this intuition, we can assume that other people act under an *expected* utility function given by

$$U(a, s) = \left(\sum_{s' \in S} R(s')T(s, a, s') \right) - C(a). \tag{14.11}$$

This equation is a simple extension of Equation 14.5, replacing the determinstic reward for an expected reward, calculated by integrating the uncertainty about the outcome that the chosen action might produce. Using this new expected utility function, we can use the same likelihood function expressed in Equation 14.6 to infer agents' preference under a probabilistic environment.

14.3 Action understanding in space and time

So far, we have focused on simple situations where agents make a single choice to reach a single outcome. In more realistic situations, action understanding involves interpreting agents that navigate in space over

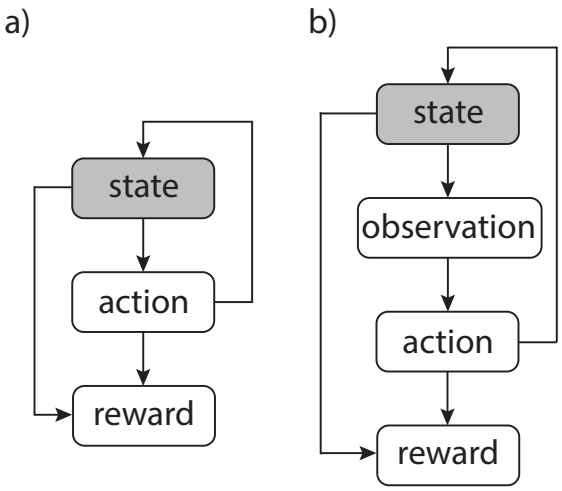

Figure 14.4: Formal models of sequential decision-making. a) Simple schematic of a Markov decision process. b) Schematic of a partially observable Markov decision process. States of the world are unobservable, but produce an observation, which guides how the agent acts (by updating their model of the world).

extended periods of time. To infer other people's beliefs and preferences in this situation, we need a model of decision-making that captures sequential planning rather than one-shot decision-making. The framework of **Markov decision processes** (MDPs), introduced in Chapter 7 and illustrated in Figure 14.4a), achieves this goal (Sutton & Barto, 2018). For simplicity, we focus on domains where the agents' task is to navigate and explore in a two-dimensional grid worlds (much like watching an agent from a bird's-eye view), as actions in these domains are sufficiently rich for people to infer beliefs, desires, emotions, and even social relations (Heider & Simmel, 1944).

MDPs represent the world as a state s from the set $\mathcal{S}$ of all possible world states. For instance, in the map shown in Figure 14.5, the world has 20 possible world states, where each state captures the agent's position in space (the world is a 5×5 grid world, but 5 positions are occupied by walls). In each state, the agent can take an action a from a set of possible actions A, such as $A = \{\text{move north}, \text{move south}, \text{move east}, \text{move west}, \text{eat}\}$.

When the agent takes an action, the state of the world changes as determined by the transition function $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$. For simplicity, suppose that agents' actions change the state of the world deterministically (although this assumption can be easily relaxed; see Section 14.2.2 above): The agent successfully moves in the intended direction unless they attempt to cross a map border or a wall (in which case they remain in the same state), and the action "eat" does not change the state of the world (but can produce a reward, depending on the state in which it's executed).

The state space, action space, and transition function specify how an agent can act in the world. Next, to capture how an agent chooses to behave, we can use a utility function $U(s, a) = R(s, a) - C(s, a)$. Notice that costs and rewards now depend both on the state and the action, which increases the expressiveness relative to the simpler utility formulation from Equation 14.5. In the context of spatial navigation, for instance, the reward can depend on taking the right action ("eat") in the right state (one where the agent has direct access to food), and costs might be different as a function of the chosen action ("eat" might have a different cost than the four physical movement actions) and the state (e.g., attempting to cross a wall could be set to incur a cost of 0 because the agent fails to move in any direction).

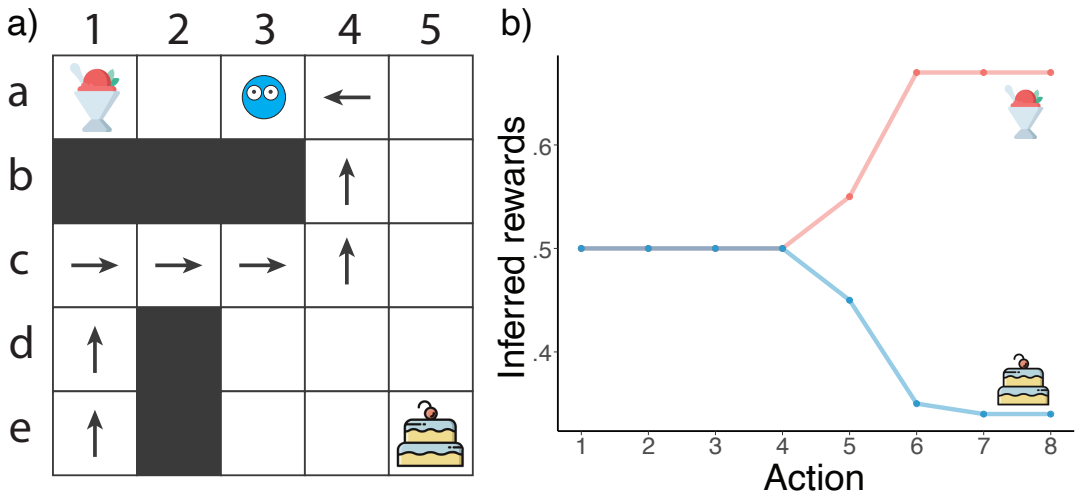

Figure 14.5: A simple example of how action sequences can be used to infer reward functions. a) Example grid-world where an agent can move in the four cardinal directions and consume the food items. b) Inferred rewards as a function of observed actionsshown in panel a.

In Chapter 7, we saw how to compute an MDP's optimal policy—a function that associates each state of the world with the action that will guarantee that the agent maximizes its utilities on the long run. This is achieved by choosing actions with the highest optimal value $V_U^*(s,a)$ given by

$$V_U^*(s,a) = U(s,a) + \max_{a' \in A} \lambda \sum_{s' \in S} T(s,a,s') V_U^*(s',a'). \tag{14.12}$$

$V_U^*(s,a)$ expresses the immediate utility $U(s,a)$ the agent obtains by taking action a in state s, plus the expected value obtained if the agent continues to take actions that maximize this value function, weighted by a future-discount parameter $\lambda \in (0,1)$. This future-discount parameter intuitively captures the idea that agents are guaranteed to obtain immediate rewards, but future expected rewards may never materialize due to unexpected events or unaccounted changes in the world.

In classical MDPs, the optimal policy is built by maximizing the value function. Thus, the optimal policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ associates each state with the action that maximized the value function $V_U^*(s,a)$. From an action understanding standpoint, however, we need to account for planning errors. Agents can occasionally make sub-optimal choices due to mistakes or accidents. This can be accounted for by softmaxing an MDP's value function to obtain a probabilistic policy, such that

$$\pi_U(a|s) \propto \exp(\beta V_U^*(s,a)). \tag{14.13}$$

Note that in MDPs, agents' actions depend only on the current state. Therefore, given an observed trajectory $t = (\mathbf{s}, \mathbf{a})$—an ordered sequence of pairs of states $\mathbf{s}$ and actions $\mathbf{a}$—the probability that the action would take each action in the corresponding state is given by

$$p(\vec{a}|U,\vec{s}) = \prod_{i=1}^{|\vec{a}|} \pi_U(a_i|s_i), \tag{14.14}$$

Thus, given an observed trajectory $t = (\vec{s},\vec{a})$, the posterior distribution over utility functions is given by

$$p(U|t) \propto \left[\prod_{i=1}^{|\vec{a}|} \pi_U(a_i|s_i)\right] p(U) \tag{14.15}$$

Figure 14.5b shows an example inference of this model. Here, an agent begins on the bottom left part of a grid world and can move in any cardinal direction. The map contains two walls—one spanning $(b, 1-3)$ and a second one spanning $(d-e, 2)$—and two sources of rewards—ice cream on state $(a, 1)$ and cake on state $(e, 5)$. For simplicity, we can assume that all actions, regardless of the state they're executed in, incur a cost of 1 (formally, $C(a, s) = 1, \forall(a, s) \in A \times S$), and that rewards are always 0, except when action 'eat' is taken in positions $(a, 1)$ or $(e, 5)$.

Figure 14.5 shows the inferred reward associated with eating cake and ice cream as a function of how many actions have been observed. The first four actions do not reveal the agent's rewards because the agent would behave the same way regardless of their preference. The fifth action—moving east—is still consistent with pursuing the cake or the ice cream. Yet, the model begins to infer that the agent prefers ice cream. This is because, in state $(c, 3)$, an agent with a reward function where R(ice cream) $>$ R(cake) will always take action "move east." By contrast, an agent with a reward function where R(ice cream) $<$ R(cake) should be equally likely to "move east" or to "move south." Finally, when the agent takes their sixth action—move north—the action is only probable when R(ice cream) $>$ R(cake), allowing the model to infer a preference for ice cream.

This approach to modeling human goal inference has been validated experimentally by several behavioral studies (Baker et al., 2009), which presented scenarios like those in Figure 14.6. Participants were asked to make goal inferences at several points along the paths of agents navigating around obstacles toward one of several marked locations (see caption for details). The conditions in Figures. 14.6(a), (b) and (c) show the same agent path, but differ in the presence of a gap in the obstacle (a), or the location of one goal object (c). These slight differences in the environment have large effects on people's goal inferences: In Figure 14.6(a), after just three steps, people immediately infer that goal A is much more likely than B or C. Figs. 14.6(b) and (c) increase the ambiguity, with goals A and B assigned similar probability in (b), and goals A, B, and C rated similarly in (c), until the agent approaches goal A after 11 steps.

The same framework be easily extended to break down inferred utility functions into the underlying costs and rewards. To achieve this, it is only necessary to treat the cost function as unobservable and variable across agents and terrain types, using Equation 14.7 with an MDP as the generative model. Consider the event shown in Figure 14.7. If we assume that actions in the same terrain must have the same cost, then cost inferences reduces to inferring two cost values and two reward values using Bayesian inference, with the likelihood term computed through Equation 14.14. Figure 14.7 shows the inferred expected costs and rewards as a function of the observed actions.

14.3.1 Distinguishing between decision making and action planning

MDPs provide a normative solution when multiple sources of costs and rewards are at play. But this formulation implicitly blends decision-making (which rewards will the agent attempt to collect?) with action planning (what actions does the agent need to take to collect them?). Yet, action planning is a hierarchical process, where agents must first choose a goal (decision making) and then take the actions to pursue it (action planning). Blending these two processes into a single computation (Equation 14.13) also limits us in our ability to distinguish between sub-optimal choices, and sub-optimal action planning: Softmax models with high noise (when β is low) assume that agents make poor choices and take poor actions, while softmax models with low noise (when β is high) assume that agents make optimal choices and take optimal actions. Yet, intuitively, these processes are dissociable and they might be subject to different degrees of sub-optimality (see Figure 14.8).

To distinguish between decision making and action planning, we can build a hierarchical model where a utility-based model identifies which goals to pursue (described in equations 14.5-14.7; Section 14.1), and an MDP computes the action plan that will fulfill each goal. More formally, consider an event with

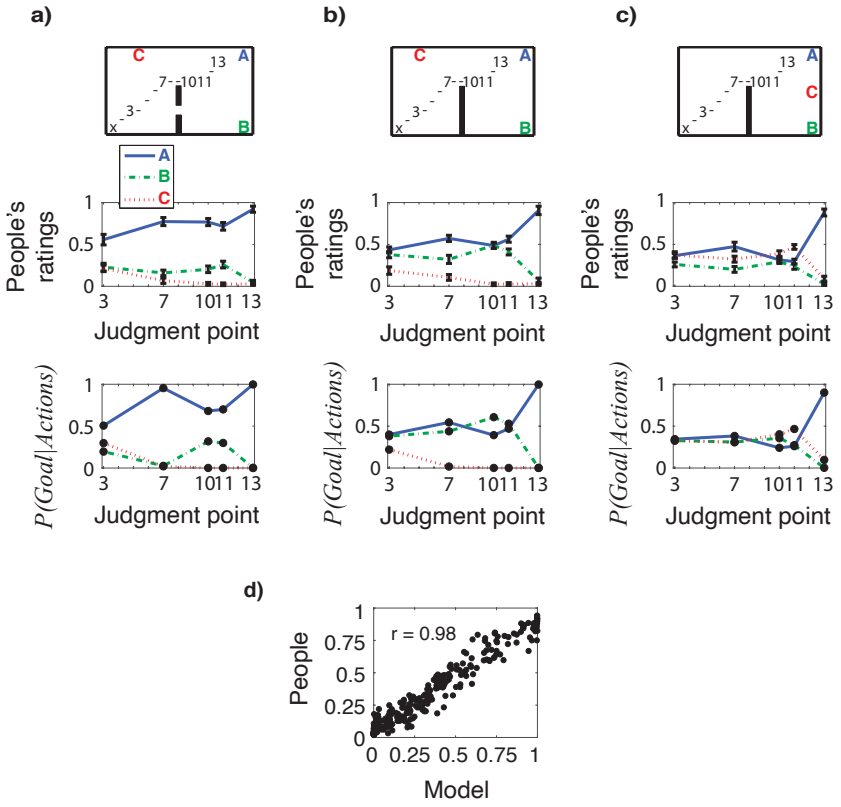

Figure 14.6: Behavioral experiment comparing human goal inferences with the predictions of a Bayesian model based on the principle of rationality (Baker et al., 2009). (a)-(c) Comparing human and model inferences in three conditions of the experiment. The top row shows a map with walls (in black), three goals (labelled A, B, and C), and the agent's trajectory. The agent's starting point is marked with an X and the numbers indicated time points where participants were asked to rate the probability that the agent was pursuing each of the three goals. The second row shows people's ratings for each goal as a function of time point in the path, and the third row shows the model predictions. (d) Quantitative comparison of human and model inferences across all conditions of the experiment.

n sources of reward, each in a different position in space (i.e., a physical environment with n objects scattered around). Let $\mathcal{G}$ be the set of possible goals—defined as every state where at least one action can yield a positive reward (that is, state s is a possible goal if $R(s,a) > 0$ for some action a)—and U_g the utility the agent obtains when pursuing goal g (as determined by Equation 14.5). The agent's probability of selecting goal g is then given by

$$p(g|U) \propto \exp\left(\beta_D U_g\right) \tag{14.16}$$

where β_D is the softmax parameter that regulates the agent's ability to select the highest-utility goal. The utility U_g associated with each goal is given by the reward associated with the final state minus the expected cost for reaching it. Note that this implies that each goal's cost will depend on the agent's initial position and the goal's location. We can calculate this cost by solving an MDP to maximize the goal-specific reward function R_g, defined as

$$R_g(s,a) = \begin{cases} R(s,a), & \text{if } s = g \\ 0, & \text{otherwise} \end{cases} \tag{14.17}$$

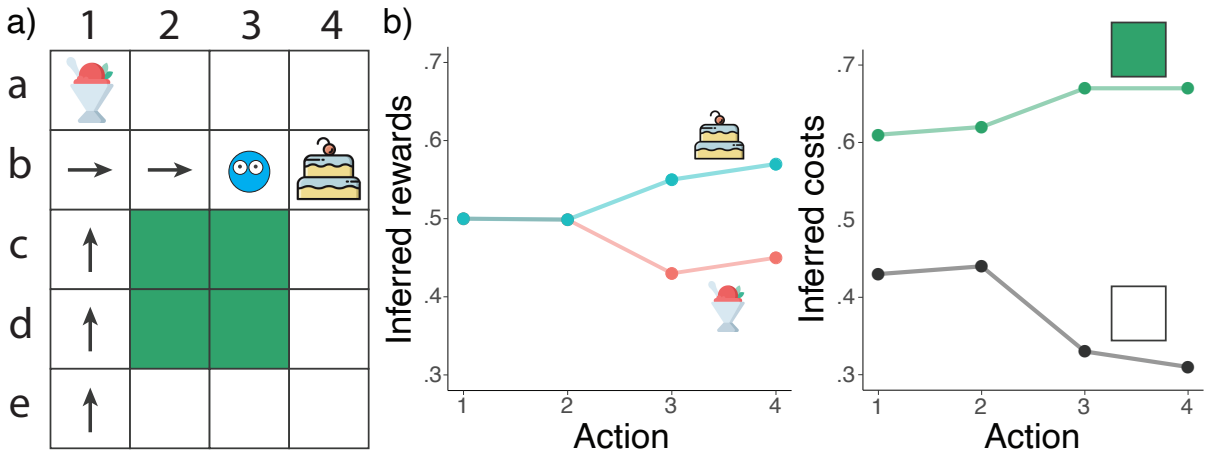

Figure 14.7: Inferring costs as well as rewards. a) Example grid-world where it is possible to jointly infer the agent's costs associated with walking through each terrain and the agent's reward associated with each food item. b) Joint inferences over costs and rewards as a function of time.

That is, this goal-specific reward function sets all rewards to 0 for any possible state in the environment, with the exception of the state identified in goal g. This reward function R_g, combined with the model's general cost function that determines the cost of different actions, allows the MDP to generate an action plan that can be used to estimate the cost of fulfilling goal g. Critically, building the probabilistic action policy π_g also involved softmaxing the action plan (Equation 14.13), which can be done using a separate softmax parameter β_A that captures the agent's ability to navigate efficiently towards its goals.

Note that because the MDP policies are probabilistic, the exact cost will depend on whether the agent makes any errors during planning. Therefore, the utility associated with each goal U_g uses expected cost

$$C_g = \sum_{t \in \mathcal{T}} C(t)p(t) \tag{14.18}$$

where $\mathcal{T}$ is the set of all possible trajectories $t = (\mathbf{a}, \mathbf{s})$ that reach the goal g from the agent's starting point. $C(t)$ is the trajectory's cost, given by

$$C(t) = \sum_{i=1}^{|t|} C(a_i, s_i), \tag{14.19}$$

and $p(t)$ is the probability that trajectory t happens, given by

$$p(t) = \prod_{i=1}^{|t|} \pi_g(a_i|s_i)T(s_i, a_i, s_{i+1}) \tag{14.20}$$

where $\pi_g(a_i|s_i)$ is the probability that the agent takes action a_i in state s_i and $T(s_i, a_i, s_{i+1})$ is the probability that this action in that state transitions the world to the next state in the trajectory. Naturally, considering every possible trajectory is intractable, but Equation 14.20 can be approximated through sampling-based methods.

From an action understanding standpoint, inference over this model requires integrating over the unobservable goal the agent selected and is acting towards. Thus, given an observed trajectory $t = (\vec{a}, \vec{s})$,

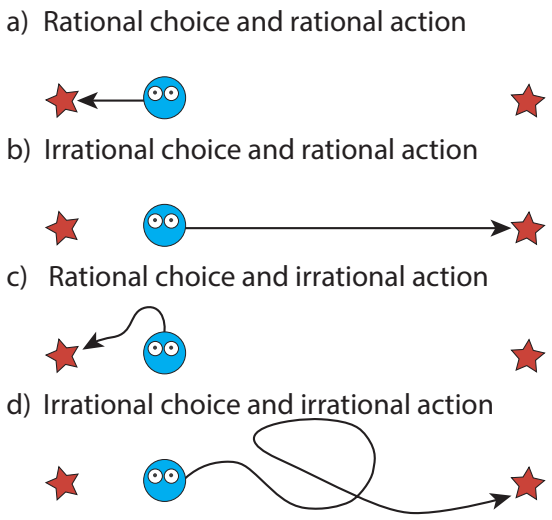

Figure 14.8: Four types of actions that distinguish between rational choice and rational action.

the likelihood is given by

$$p(t|U) = \sum_{g \in G} p(t|g)p(g|U) \tag{14.21}$$

where $p(t|g)$ is the action-planning model given a target goal (Equation 14.13 with reward function 14.17), and $p(g|U)$ is the decision-making model given the set of rewards (Equation 14.16). Finally, this likelihood, weighted by a prior over utility functions $p(U)$ yields the posterior distribution over agents' underlying utilities.

Jara-Ettinger, Schulz, and Tenenbaum (2020) showed how this model captures people's capacity to jointly infer other people's costs and rewards based on how they act, using scenarios like the one shown in Figure 14.9a. Here, an agent must travel from a starting point (middle left) to a target location (middle right), but has the option of collecting two objects on the way (a white cube or an orange cylinder). The agent's path immediately reveals that navigating through the blue terrain is less costly than navigating through the purple terrain (otherwise, why take the longer path?). Both the model and people infer that the agent does not like the orange container because the agent could have gotten it by taking an equally costly path. By contrast, both the model and people are more uncertain about the white box. Although the agent chose not to get that box, it's also in the middle of a region that we inferred was costly to go through, making it possible that the agent liked the box but chose not to pursue it due to the costs involved (Figure 14.9b). As Figure 14.9c shows, this model captures people's inferences in a broad range of events.

14.3.2 Planning under uncertainty

Agents often face situations where they do not know the exact state of the world or the position of different target goals. In these situations, Markov decision processes are no longer appropriate, as they assume perfect information. An extension of MDPs, called **partially-observable Markov decision processes** (POMDPs) helps model agents' behavior in these contexts.

To model planning under uncertainty, we first need to expand the state space to include world states that the agent may consider plausible, even if they cannot ever occur. In the example in Figure 14.5,

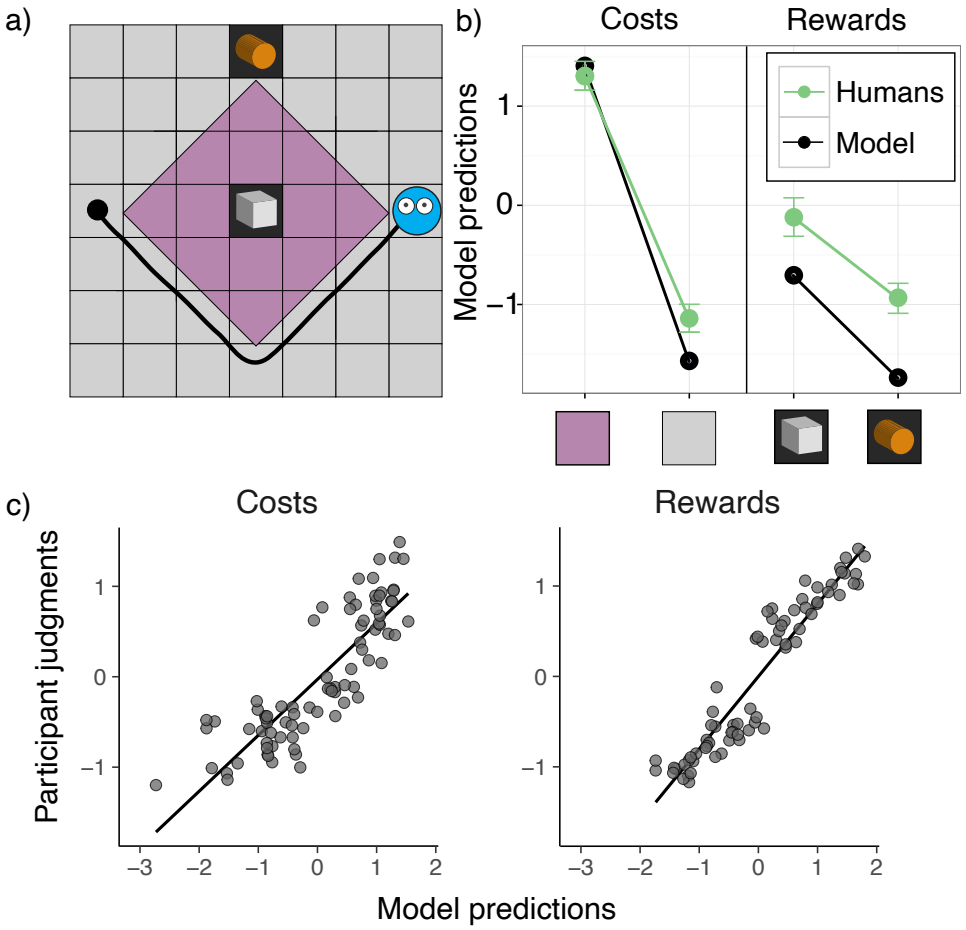

Figure 14.9: Results from joint inferences over costs and rewards. (a) Example scenario: An agent navigates from a starting location to a target location. The map contains different types of terrains, and objects that the agent can collect if they desire. (b) People's inferences about the underlying costs and rewards, and model predictions for the example in panel a. (c) Overall comparison between human joint cost-reward inferences and model inferences.

the state space consisted of 20 possible states, each capturing the agent's position in space. To model an agent who may not know whether the ice cream is in the top left position or in the bottom right position, we would need to expand the state space to 40 states, with each state capturing the agent's position in space and the position of the ice cream and the cake. In this extended state space, the world is always in one of the original 20 states (where the ice cream is always on the top left), but the agent may believe they are in a world state that does not match reality (e.g. believing that the cake is in the top left position). Under this expanded state space, we can define an agent's beliefs $B : \mathcal{S} \rightarrow [0, 1]$ as a probability distribution over states of the world.

When agents act under their beliefs about the state of the world (instead of acting based on the true state of the world), the policy must now map beliefs onto actions (rather than states onto actions). Formally, we require a policy $\pi_U : \mathcal{B} \times \mathcal{A} \rightarrow [0, 1]$ such that $\int_{a \in A} \pi_U(b, a) = 1$ for any belief b and any utility function U (i.e., for any belief, the probabilities over all actions must add up 1, ensuring they express a proper probability distribution).

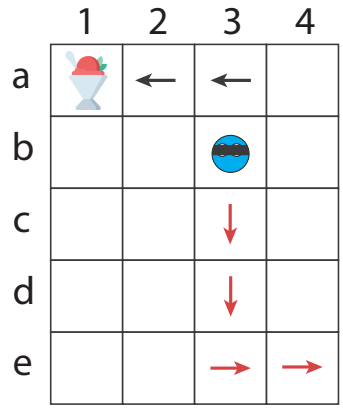

Figure 14.10: Example of how an agent with partial knowledge about the world can take actions to infer its location and obtain rewards. Here, an agent wearing a blindfold is initially unsure about their position in space. If they can detect walls, then the agent can infer their location in space by reaching the northwest (black arrows) or southeast (red arrows) corners. Reaching the northwest corner, however, is a better strategy because it also guarantees that the agent will reach their goal state.

In addition, we need to specify how agents' beliefs change as they interact with the world. To achieve this, POMDPs assume that each combination of states and actions produce observations about the world that the agent can use to make inferences about the true world state. For instance, an agent walking into a room may receive an observation that reveals what is inside.

Formally, let Ω be the set of possible observations that the agent can receive (i.e., the full set of all possible information that the agent could get as they interact with the world). In a POMDP, the agent receives one observation $o \in \Omega$ at each time step, determined by an observation function $O : \mathcal{A} \times \mathcal{S} \times \Omega \rightarrow [0, 1]$ where $O(a, s, o)$ is the probability of getting observation o in state s after taking action a. As the agent receives observations, they update their beliefs through

$$p(s|a, o; b) = \sum_{s_o \in S} T(s_o, a, s)O(a, s, o)b(s_o). \tag{14.22}$$

The left-hand side represents the agent's belief that they're in state s after taking action a and receiving observation o, given that they previously had beliefs b. The right-hand side computes this term by considering all of the possible states that the agent might have previously been in ($s_o \in \mathcal{S}$). For each potential previous state, $b(s_o)$ is the agent's belief that they were in such state, $T(s_o, a, s)$ is the probability that action a in state s_o would transition to state s, and $O(a, s, o)$ is the probability of receiving observation o after taking action a to reach state s.

To illustrate the dynamics in a POMDP, consider an agent with a blindfold moving in a 5×4 grid world (Figure 14.10). The state space S consists of 20 states, each indicating the agent's position in space. Suppose that the agent cannot see anything, but they can feel the walls whenever they try to cross one (and hence hit it). The space of observations is then $\Omega = \{\emptyset, \text{wall}\}$. Any action-state pair within the map would produce observation $\emptyset$, and any action in a state where the agent hits a wall produces observation "wall." An agent with complete uncertainty ($b(s) = 1/16$ for all $s \in S$) can determine their position in space by moving south until getting the observation "wall" (at which point they will know they must be in one of the bottom four states) and then moving east until getting the observation "wall" again (red path in Figure 14.10). At this point, the agent will know that they must be in the bottom-right corner. If actions change the state of the world in a deterministic way, the agent will then know the exact state of the world at every time point by simply tracking which actions they took, enabling them to navigate

towards the reward. A better strategy, however, would be to move north until hitting the wall and then move west until hitting a wall again (black path in Figure 14.10). This is because the top left corner not only reveals the state of the world, but also leaves the agent in the state that has a reward. Solutions to POMDPs naturally produce policies that combine actions in the service of reducing uncertainty with actions in the service of obtaining rewards, making them a natural framework for understanding agents whose behavior is a combination of exploration and exploitation.

Under partial knowledge, we can define the utility (from the subjective point of view of the agent) of taking action a under beliefs b as

$$U(b, a) = \sum_{s \in S} b(s) U(s, a) \tag{14.23}$$

and the optimal value of belief b as

$$V_U^*(b) = \max_{a \in A} \left(U(b, a) + \lambda \sum_{o \in \Omega} O(s, a, o) V_U^*(b') \right). \tag{14.24}$$

This equation is equivalent to Equation 14.12, extended to account for the agent's uncertainty. Here, the optimal value of belief b is calculated by considering the action that yields the highest utility, plus the expected future value (discounted in time by parameter λ) by integrating over the possible information that the agent might receive, and the agent's updated beliefs $b'(a, o, b)$ after taking action a, receiving observation o with initial beliefs b. These updated beliefs are calculated through Equation 14.22.

As in MDPs, we can build a probabilistic policy by softmaxing this value function (Equation 14.24). Using the resulting policy as the generative model, joint belief (the probability distribution over states), desire (the latent reward function) and competence (the underlying cost function) can be inferred through Bayesian inference, where given an observed trajectory t,

$$p(B, R, C|t) \propto p(t|B, R, C) p(B) p(R) p(C). \tag{14.25}$$

Baker, Jara-Ettinger, Saxe, and Tenenbaum (2017) developed and tested the model we have just presented, asking adult participants to make joint inferences about agents' beliefs and desires based on how they navigated an environment. Figure 14.11 shows several scenarios from this experiment, in which a hungry graduate student leaves their office to walk to lunch at one of three food trucks: Korean (K), Lebanese (L), or Mexican (M). There are two parking spots for the trucks (marked in yellow), and trucks can park in different spots on different days, or not show up at all, so the student may not know where each truck is parked, and must plan carefully where to walk in order to get lunch from the best truck available as quickly as possible. Using a POMDP for a generative model, the agent's desires can be captured using a reward function which represents their preferences over trucks, and the agent's initial beliefs can be represented as a probability distribution over each of three partially observable world states: the Northeast parking spot being occupied by (1) Lebanese (L) or (2) Mexican (M), or (3) being empty (N for none). Finally, observations of the trucks are determined by line of sight, with a small probability of observation failure.

Consider Figure 14.11c, in which the student can initially see the Korean truck in the Southwest parking spot, but cannot see the Lebanese truck parked in the Northeast parking spot due to the building blocking their view. The student walks past the Korean truck, continues walking around the building, sees that the Lebanese truck is in the Northeast spot, and then turns and walks back to eat at the Korean truck. Based on this information, which truck did they want the most? And which truck did they believe was in the Northeast spot? Participants here infer that the student wanted Mexican food most, and Lebanese food least. Participants also attribute an optimistic initial belief that the Mexican truck was in the Northeast spot.

The model captures people's inferences that Mexican is most preferred, and Lebanese least, and people's attribution of a false belief that Mexican was present. The POMDP model naturally explains the

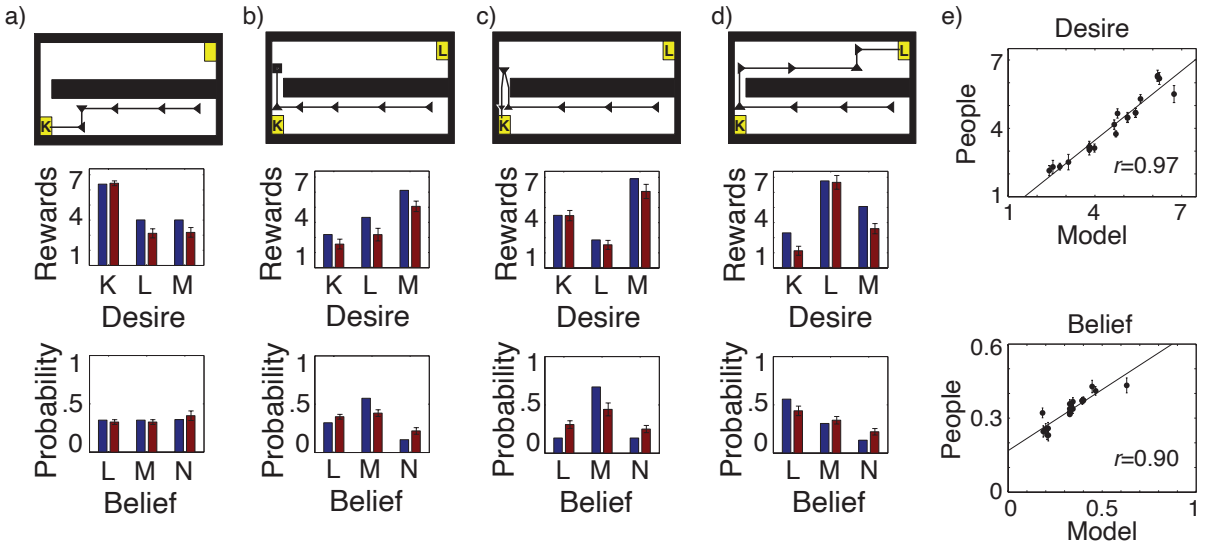

Figure 14.11: Experimental validation of Bayesian inference over beliefs and desires using a POMDP as a generative model. The experiment presented scenarios in which a hungry student seeks lunch from one of several food-trucks: Korean (K), Lebanese (L), or Mexican (M). People made different belief and desire attributions in each scenario, which the model captures with high accuracy. a-d) Comparing participant and model inferences across a range of experimental scenarios. In each panel, the top row shows the stimuli, the second row shows inferred rewards for each of the three food trucks, and the final row shows inferences about what Harold expected would be on the top right spot (with N corresponding to the belief that there was nothing there). e) Correlation between participant judgments and model inferences.

action of going around the building to check which truck is in the Northeast spot as rational exploration: Seeking information here is rational if either Mexican or Lebanese have greater reward than Korean, and if the prior belief that Lebanese or Mexican could be in the Northeast spot is non-zero. Figure 14.11b shows a case where the student paused after going around the building. Here both people and the model attribute greater belief and desire for Lebanese and Mexican (the slightly greater values for Mexican than Lebanese are due to perceiving the "pause" in the trajectory animation as hesitation on seeing Lebanese, which is interpreted as a preference for Mexican).

Once the agent turns back after observing Lebanese in the Northeast spot (Figure 14.11c), only reward functions where Mexican is preferred to Korean, and both Korean and Mexican are preferred to Lebanese, can explain the entire trajectory, and these are inferred by the model. The model's false belief inference stems from the fact that hypotheses which assign low prior probability to Mexican being in the Northeast spot are not consistent with the observed behavior; for these hypotheses, the rational action is to go straight to the Korean truck without seeking to observe the Northeast spot first. Figure 14.11 shows more scenarios in which the model captures people's judgments and Figure 14.11e shows correlations between average human desire and belief judgments and model predictions across all scenario types.

14.4 Minds thinking about themselves and other minds

So far we have considered how to infer beliefs, costs, and rewards given someone's behavior. However, mental-state inferences are often in the service of other tasks such as learning from others, deciding if someone is nice or mean, or learning what goals are worth pursuing. In this section we describe how to

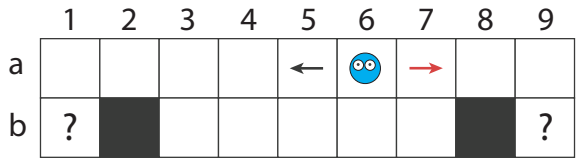

Figure 14.12: Example of how it is possible to learn about the world by watching how other agents act. A reward is located in $(b, 1)$ or $(b, 9)$. If the agent navigates left (gray arrow), the agent must believe the reward is in state $(b, 1)$. If, instead, the agent navigates right (red arrow), the agent may know the reward is located in state $(b, 9)$ or may be uncertain and hence check the spot closest to their starting location.

use Bayesian Theory of Mind models to solve these problems.

14.4.1 Tracking knowledge and learning from experts

While we have now considered cases where agents have incomplete or incorrect knowledge about the world, all of our examples always used situations where the observer (i.e., the reader, or the participant in the experiment) had perfect information about the world. In everyday social situations, we often have to interpret the actions of agents who may know more about the world than we do.

Consider Figure 14.12, which shows a grid world with an object that could be located in state $(b, 1)$ or in state $(b, 9)$. If we assume that the agent knows the object's location, then their actions will reveal this: If the agent moves left, then the reward must be in $(b, 1)$, and if the agent moves right, the reward must be in $(b, 9)$. We can formalize these intuitions by inferring the state of the world based on the agent's behavior, given by $p(s|a) \propto p(a|s)p(s)$. Here $p(s)$ is the prior probability that the world is in state s (which is known to the agent), and $p(a|s)$ is the likelihood, given by the softmaxed policy that an agent with perfect knowledge (hence modeled as an MDP) would take the observed actions a if the true state of the world were s.

In more complex cases, both the observer and the actor may have incomplete knowledge about the world. Continuing with the example above, suppose that you do not know whether the agent knows the object's location. If you watched the agent navigate to the right, reach state $(a, 9)$ and then retrace their steps, you could immediately make two inferences: (1) the agent did not know where the object was located, and (2) the object must be in position $(b, 1)$. By contrast, if the agent began by taking a step to the left, a single action would already give us some confidence that the object is in position $(b, 1)$ and that the agent knew this (otherwise, why would the agent begin searching by incurring a high cost?).

We can model these intuitions as a joint inference over agents' initial beliefs b and the true state of the world s (using a POMDP as the generative model). To achieve this, we must compute the probability that an agent with beliefs b in world state s would take actions a, $p(a|b, s)$, which can be obtained by considering how a rational agent would act under different beliefs in different world states.

Figure 14.13 shows the results from a study testing this capacity in humans (Jara-Ettinger, Baker, & Tenenbaum, 2012). Participants watched an agent navigate a simple environment (Figure 14.13a) and they had to infer the location of different food carts. The world contained three food carts (see Figure 14.13b for all arrangements), and the agent preferred A over B, and B over C. However, each food truck could be either open or closed, such that the agent couldn't always eat at his favorite truck. In the path in Figure 14.13a, for instance, we can infer that there is an open cart in the north position because the agent ultimately went to it. However, this cart cannot be cart A. Otherwise, the agent would have walked straight there rather than checking other hallways. The fact that the agent never checked the west aisle

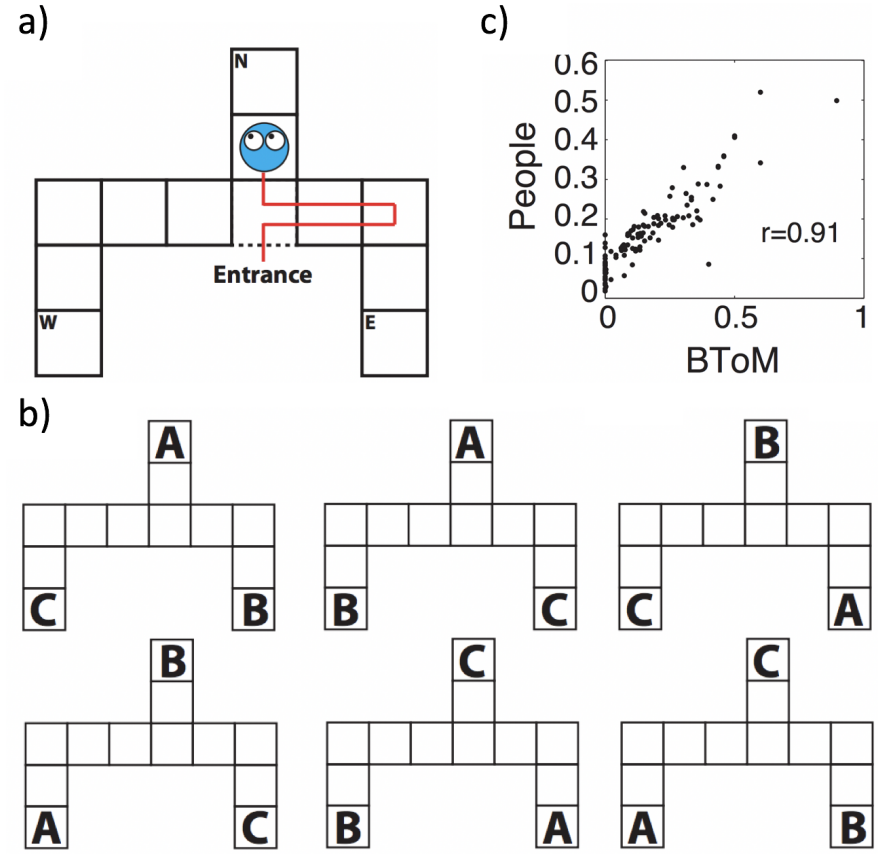

Figure 14.13: A more complex example of learning about the world via another agent's actions. a) Example trial of Experiment 2 in Baker et al. (2017). An agent navigates a set of hallways with three food carts (positioned in N, W, and E) to decide where to eat. The agent always prefers A over B and B over C, but these food carts are sometimes closed. b) Possible world states. Participants had to rate the probability of each world given how the agent behaved. c) Overall correlation between our model and participant judgments.

suggest that cart A is also not there, otherwise the agent would have gone there to check if it was open. Together, this suggests that the agent first saw cart B open in the north spot, then found cart A closed in the east spot, inferred that cart C must be in the west spot, and therefore went back to eat at cart B. Panel c shows how this computational model tracked participant inferences with quantitative accuracy.

14.4.2 Helping and hindering

The rewards and costs described so far were associated with concrete world states and actions: Does my friend want cake, or ice cream? how much do they cost? and so on. But people also care about other people. People's goals, costs, rewards, and intentions can take into account and be directly related to the goals, costs, rewards, and intentions of others. This is true both for our own planning and decision making, and in our inference over mental states. To capture this kind of reasoning, we need to extend our framework to include multi-agent planning, and allow beliefs and utilities to operate over other mental variables.

To continue with our running example, if my friend forgot their wallet and wants to get some ice cream, I may order one for them. My preference is not for them to have ice cream, per se, but for my friend to have what they want. That's part of what being a friend is. And if I see a person paying for a person's ice cream, I might reasonably conclude they are friends. Even young children seem capable of such reasoning (Hamlin, Wynn, & Bloom, 2007; Hamlin & Wynn, 2011; Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013). How can we formalize this commonsense understanding in our framework?

Let us consider a simple case in which one agent, A, has a utility function U^A that maps every combination of states and actions into the difference between the costs and rewards. Meanwhile a second agent, B, has a utility function U^B that is a function of agent's A utility:

$$U^B(s, a^B) = \rho[U^A(s, a^B)] - c(a^B), \tag{14.26}$$

Here, agent B's utility for taking action a^B in state s is given by a constant $\rho < 0$ multipled by the expected utility that agent A will receive when that action is taken, minus the cost that agent B incurs to take this action. Note that in many cases, agent B might not know agent A's exact utility, in which case the first term of the right-hand side can be replaced with the expected utility rather than an exact one. ρ is a parameter that controls the direction and degree of the social preference—how much B "cares" about A. When $\rho > 0$, B will help A, by taking actions to achieve states that are favorable to A. When $\rho < 0$, B will hinder A, by doing the opposite. Both of these are balanced by the cost: If I'm your friend I'll buy you ice cream, but I'm not giving you my car. Given an agent with such a utility function, social reasoning about goals can be turned into a straightforward question of inference over the ρ parameter.

As before with joint reasoning about costs and rewards, if we do not know another agent's ρ nor their cost, their specific actions could be explained in different ways. An agent's lack of willingness to help could be explained by the cost being too high, or the motivation (ρ) not being high enough. If an agent refuses to help when the cost is known to be low (e.g. the agent is adjacent to the known reward and easily hand it to the other agent), the model will infer a lower ρ than when an agent's costs is high (e.g. if the social agent is more distant to the inferred reward than the first agent, such that even if they tried to help, they would get to the reward long after the first agent did).

The inference problem remains largely the same as that described for most of this chapter: given a set of actions, infer the underlying utility (rewards, costs), beliefs, intentions, and now we add the parameter weighting the utility of others. As before, the likelihood for actions is given by a model of planning. In a social setting, this means a multi-agent planning model, with possible uncertainties and partial observability depending on the assumptions about the beliefs of the agents involved.

It is worth pausing here to consider an alternative account of social goal inference, one that does not rely on utility functions that operate over utility functions. Social goal inference has also been cast as a problem of relating perceptual cues (in particular visual cues) directly to inferences over mental states. Certainly some actions seem to relate perception directly to mental inference. Slapping a person means you're not friends with them; why go through the computationally expensive calculation of inverting a planning procedure to figure that out? If someone asked for ice cream and you gave it to them, you're their friend. Simple. Various accounts have tried to formalize such cue-based social inference. For example Barrett, Todd, Miller, and Blythe (2005) related a host of motion patterns directly to the inference of social intentions such as 'guarding' or 'courting'.

On the face of it, cue-based accounts are simpler than reasoning about probabilistic planning. However, they also require learning and training many cues for many situations, and they have trouble generalizing. Consider that the ρ parameter is not tied down to a particular state or action. In a specific situation, helping another person may require getting close to them. Other times helping may require stepping away. The same action could be interpreted as harmful or helpful depending on the mental states of the agents involved.

To test this idea, Ullman et al. (2009) contrasted an inverse-planning model of social goal inference to a cue-based one, in a domain similar to that used by Baker et al. (2009). In this domain, two agents

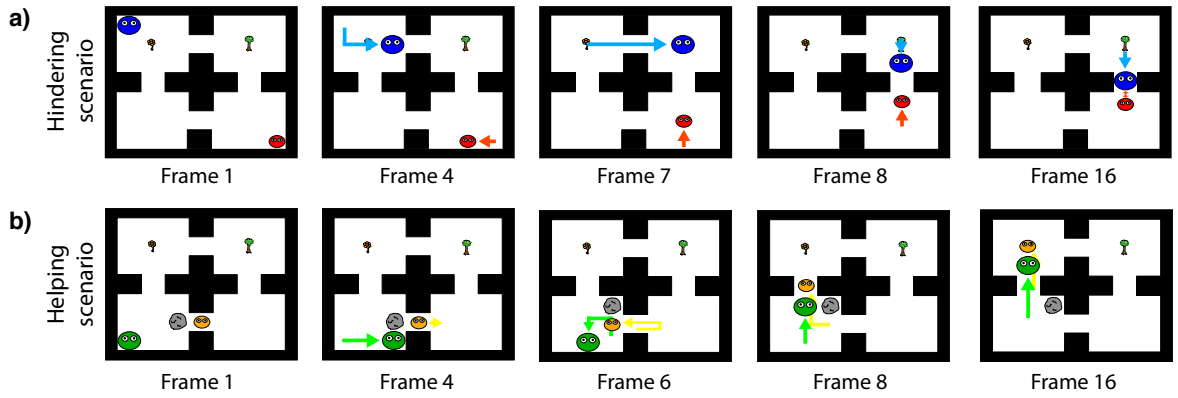

Figure 14.14: Still frames of example behavior sequences used in the helping and hindering domain used in Ullman et al. (2009). Both agents begin in frame 1 and progress in discrete steps, as shown by the colored arrows. The sequences were paused at different probe points and participants were asked to infer the agents' goals.

pursued selfish and social goals in a small maze environment (see Figure 14.14). The small agent was always selfish, in that it was trying to get to one of two goal points (a flower or tree). The large agent was either selfish or social. Social agents try to help or hinder the little agent. The small agent occasionally failed in its action, mimicking the behavior of agents in Hamlin et al. (2007). Large agents never failed in their action, and could push small agents around. The environment occasionally included a boulder which only large agents could move. Different scenarios used different initial locations of the goals, agents, and boulder, and different goals for the agents.

Participants were asked to judge the goals of the large agent at different time points and their judgments were compared to a Bayesian Theory of Mind model that included multi-agent planning, and to a cue-based model that included 10 different visual cues from previous applicable cue-based models, including cues such as geodesic distance, changes in distance, and relative movement. As shown in Figure 14.15, people's judgments about the large agent's goal included moments of certainly and confusion, rapid switches, and growth in confidence. The cue-based model was able to recover the correct goal when the goal was selfish (using cues such as shrinking distance to the tree or flower), but it was far less accurate about social goals. The cue-based model was also poor at generalizing: When trained on scenarios that involved boulders and tested on non-boulder scenarios or vice-versa, its performance suffered greatly. By contrast, the planning-based model was able to recover human judgments with high precision, its accuracy was unaffected by whether the goal was selfish or social, and it performed equally well on boulder and non-boulder scenarios.

While the simple toy-domain shown in Figure 14.14 is a useful arena for pitting different accounts against one another, it is also far simpler than real-world situations. As the most basic extension, the situation can be made more complicated when we consider that agents could be balancing social rewards with their own rewards (I have other things going on besides your ice cream), but this is a straightforward extension to Equation 14.26. Of course, situations in which multiple agents are reasoning and acting with regards to the social goals of others (while being uncertain about the rewards and costs of others) are complicated. One could quickly end up with situations in which the model would reason that A might be trying to help B even though A knows that B doesn't want A's help, but A thinks that's only because B believes that A doesn't understand what B's goal is, and if A could convince B, and so on, and so on.

Another drastic over-simplification is that this models required fully specifying the (small) space of possible goals. This led on occasion to odd situations in which people's understanding of social behavior

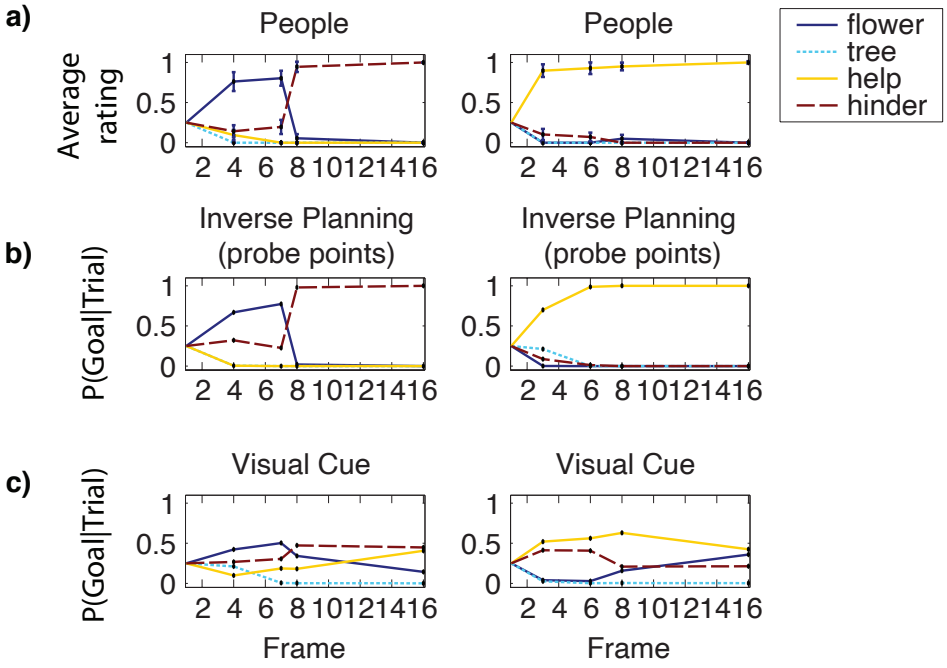

Figure 14.15: Example human judgments compared to model predictions for the task used in Ullman et al. (2009). The first column corresponds to the hindering scenario shown in Figure 14.14a and the second column corresponds to the helping scenario shown in Figure 14.14b. a) Average participant judgements about the large agent's goal at different probe points. Horizontal black lines show standard error bars. b) Predictions of Bayesian Theory of Mind model at probe points. c) Predictions of cue-based model used in Ullman et al. (2009).

and the planning-based model diverged. For instance, consider a situation in which the small agents struggles and struggles and eventually reaches a flower, while the large agent simply hangs out in a far off corner. What is its goal? The model used in Ullman et al. (2009) quickly identifies this as "helping." The logic is the following: If the large agent was trying to get to the flower or tree, it would move towards them. If it was trying to hinder, it would move to hinder. Since it did none of these, it must be that the large agent knew that by the time it reached the small agent, the small agent would already have reached the goal. Since movement is costly, why bother budging? People clearly did not share this intuition. This does not suggest that the planning-based modeling approach is wrong, but rather that this simple instantiation is too simple. A fuller model would include a richer space of goals, beliefs, communication (e.g. moving slightly towards the other agent to indicate willingness to help while not committing). But these are all additional cogs in the basic machinery, not a new machine altogether.

14.4.3 Honesty and reputation

In social situations like the ones we presented above, inferences about helping or hindering were achieved by allowing agents to have desires over other people's desires. The same approach can be used to formalize the idea that agents can also have desires over other people's beliefs.

People's desires over other people's beliefs captures a range of types of social behavior that are common, like inferring that an agent is motivated to share their knowledge, or that they want the agent to have an incorrect belief (i.e., lying). When we consider that agents have beliefs about each other, this

approach can help model people's reputational concerns. We can achieve this by building a meta reward function over beliefs $\text{MR} : \mathcal{B} \to \mathbb{R}$ where $\text{MR}(b^B)$ is the reward an agent gets when agent B's beliefs are b^B, and $\mathcal{B}$ is the space of possible beliefs. In many cases, agents do not care about the full set of beliefs that an agent has, and only care about some dimension. For instance, A might want to ensure that B believes that A is helpful, without caring about B's remaining beliefs about the world or about others. These types of situations can be captured by associating a reward different critical features of others' beliefs, and computing MR as the sum of these features.

Given a personal reward function R_p and a meta reward function MR, we can define the final reward function R_f as a function that maps states of the world and observer beliefs onto rewards through

$$R_f(s, b^A) = R_p(s) + \theta \int_{b^B \in \mathbb{B}} \text{MR}(b) b^A(b^B)\, db^B \tag{14.27}$$

where ρ is once again a parameter indicating how much the agent cares about their reputation, and $b^A(b^B)$ is agent A's belief that B has belief b^B.

14.4.4 Effort and motivation

Finally, the models developed so far assumed that the relation between states and actions is given by a transition function where the probability of switching from state s to state s' after taking action a is constant, determined by $T(s, a, s')$. More realistically, agents can exert different degrees of effort that affect the probability that their actions will be successful. We can expand past frameworks by assuming that agents' behavior is a combination of an action and some effort $e \in [0, 1]$ that affects the action's cost and its probability of success. Given an action space $\mathcal{A}$, we can define the effort-based action space $\text{CA} = (\mathcal{A} \times [0, 1])$ and an effort-based cost function $C : \mathcal{A} \times [0, 1] \to \mathbb{R}$ can be expanded to take effort into account, such that

$$C(a, e) = C(a)^{(c+e)}. \tag{14.28}$$

Here, the cost of an action is expontentiated by a constant value c and the amount of effort e put into the action.

Similarly, the transition function must now account for the amount of effort the agent placed, which can be achieved by building a transition function that takes effort as a parameter and concentrates the distribution of expected outcomes accordingly. The exact nature of this function, however, depends on the event. In some cases, low-effort may lead to failure and cause the agent to remain in their previous state, in other cases, low-effort may cause unintended negative consequences and leave the agent in a less favorable state than their original one. Note that because the action space and the transition function are continuous, these problems must be solved in a non-discrete framework, such as continuous MDPs (Sutton & Barto, 2018).

14.5 Future directions

The framework presented above captures the core computations behind our ability to make sense of other people's behavior. In the remainder of this chapter, we discuss what's missing, and how this framework provides a starting point to tackle questions that any complete theory of human social cognition must explain: How can we capture human reasoning where goals are richer or more abstract than simple navigation towards objects? How do we represent and reason about different types of minds? How do these mental-state inferences work in more realistic situations where we see bodies moving in three dimensions rather than points in a two-dimensional world? And how are these computations implemented in the brain?

14.5.1 Inferences over types of goals and types of minds

Theory of Mind models and tasks have historically focused on a limited class of goals: reaching for objects or navigating towards different positions in space (and helping or preventing others from reaching certain objects or navigating towards a position in space). The space of goals that people pursue, however, is much broader, including purely epistemic goals (learning something for the sake of learning it), communicative goals (moving with the goal of sharing a message, such as when we gesture), and even creative goals like improvising tools for ad-hoc tasks.

Intuitively, these complex goals can be expressed as compositions of simpler ones that involve moving in space to elicit rewards that can be obtained from a relatively abstract space (e.g., rewards for obtaining information or rewards for revealing communicative intent). One way to expand these models is by building hierarchical goals where more abstract goals are fulfilled by performing sequences of basic types of actions. Under this view, recognizing abstract goals requires parsing other people's behavior in terms of simple actions, and then having access to a space of abstract goals that can be fulfilled through those actions. In recent work, Velez-Ginorio, Siegel, Tenenbaum, and Jara-Ettinger (2017) showed how a space of unbounded goals can be formalized within a probabilistic context-free grammar that combines primitive goals to represent richer ones. In this approach, desires are not longer represented as rewards. Instead, they are represented as propositions which, when true, elicit a reward. Inferring an agent's desire is then a process of inferring what type of proposition an agent was attempting to make true.

More broadly, people are not only able to reason about different types of goals, but also about different types of minds. Intuitively, the behavior of a human adult is driven by more complex psychological machinery than the behavior of a toddler, and the psychological machinery behind the behavior of a toddler is incommensurable relative to the psychological machinery of non-human mammals. These intuitive differences suggest the beliefs, desires, and cognitive processes that we attribute to others have different representational scope (e.g., while we think of humans and hamsters as both capable of having beliefs about the existence of an object, we might intuitively believe that only humans can have beliefs about its historical significance). Even in simple situations where we can assume that all agents have similar beliefs and desires, we nonetheless expect different types of agents to create plans with different degrees of complexity. Recent work has found initial evidence that inferences about types of minds can be expressed as Bayesian inference of a space of planning and decision-models of varying complexity (Burger & Jara-Ettinger, 2020). More research is needed in characterizing the variability in the different types of minds that we can represent, and formalizing this space to explain how people can jointly infer another agent's type of mind as well as their mental states.

14.5.2 Mental-state reasoning from visual input

The models presented here operate over simple spaces of actions in two-dimensional displays. A key challenge lies in extending these models to full-body control. Modeling moving bodies rather than dots moving in simple two-dimensional space not only expands the space of situations that these models can reason about, but also provides a more natural framework for understanding additional inferences that people make in social contexts. For instance, as observers, when watching another agent pursue a goal, we may want to estimate its difficulty to decide whether we should engage in it ourselves or not. Recent work shows that these problems can be solved by reasoning about motion planning. Given a set of arranged blocks (e.g., blocks sorted by colors, or arranged into a tower), Yildirim et al. (2019) showed that difficulty can be estimated by calculating the energy that an agent would have to expend to build the tower (through an ability to simulate full-body control), and the risk of the construction, given by the chance that the arrangement might collapse.

A further challenge is to model different degrees of granularity depending on the task at hand. When we reason about agents moving around in a city, we intuitively conceptualize them in a similar way to

how the models above handle two-dimensional motion. When reasoning about events in smaller spatio-temporal scales, like an agent moving in a room and interacting with objects, however, we need to model their arm movements in order to interpret reaching and grasping events. And when reasoning about more fine-grained situations, like an agent manipulating a tool, we require an even more precise model of hand control. Thus, a critical challenge lies in building models of action-understanding that can flexibly switch between levels of granularity depending on the goal that the agent is pursuing.

14.5.3 Neural substrate of inverse planning

Finally, our Bayesian models are best conceptualized at a computational level of analysis, and we know little about their underlying algorithmic implementation. One outstanding challenge is to explain how these models may be implemented in the brain. Broadly, two types of solutions that have been used in similar problems can serve as a starting point.

A first possibility is that complex computations are implemented as a collection of simple heuristic that, combined, approximate normative inferences. If so, cheap rules that exploit contingencies between low-level visual features and the corresponding mental-state inferences may allow people to avoid doing any kind of Bayesian inference, while still obtaining some approximate solution (although note that previous attempt at this have been unsuccessful; see Section 14.4.2). One challenge to this idea, however, is that people's mental-state inferences are not only qualitatively, but also quantitatively, consistent with Bayesian models, and a solution of this type would have to explain how a collection of rules may approximate these models with such quantitative accuracy.

A second possibility is that Bayesian inference is "compiled" into sub-symbolic computations that allow these inferences to be performed in more of a feed-forward fashion than is typical for Bayesian inference algorithms work, similar to how face recognition implements Bayesian computations (Yildirim, Belledonne, Freiwald, & Tenenbaum, 2020). It is also possible that only some aspects of action-understanding can be implemented in a feed-forward fashion, perhaps explaining the boundary between the perceptual component of how we detect agents and goals, and the more cognitive component of inferring beliefs and desires (Scholl & Tremoulet, 2000).

14.6 Conclusion

In 1944, psychologists Fritz Heider and Marianne Simmel published "An Experimental Study of Apparent Behavior" (Heider & Simmel, 1944), where they showed how a simple animation of geometrical shapes moving in a two-dimensional space can elicit rich social inferences. People watching this video for the first time often report seeing these shapes spontaneously come to life, and inferences about the shapes' goals, beliefs, desires, and intentions are so strong that they almost feel *visible*. The framework we presented here—expressing actions in terms of abstract mental states in a hierarchical model that captures how minds relate to behavior—provides a theoretical foundation for explaining how social reasoning often goes beyond the data, allowing us to determine what others think and want, what they think they want, how they feel towards other people, and even what they think about themselves.

At the same time, watching Heider and Simmel's classical animation also reveals the long road ahead. A few seconds of this video not only reveal each individual agent's mental states; they also reveal the nature of their social relations—who is brave, who is a bully, and which agents are friends and which are not, to name only a few. The work presented here, we hope, will be foundation for building richer and more powerful models that not only capture inferences about individual minds, but also inferences about the complex social actions that compose social relationships.

References

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*, 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, *26*(4), 313–331.

Burger, L., & Jara-Ettinger, J. (2020). Mental inference: Mind perception as Bayesian model selection. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society.*

Csibra, G., Bíró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, *27*(1), 111–133.

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, *26*(1), 30 – 39.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*, 557–559.

Hamlin, K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental Science*, *16*(2), 209–226.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, *57*(2), 243–259.

Jara-Ettinger, J., Baker, C., & Tenenbaum, J. (2012). Learning what is where from social observations. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society.*

Jara-Ettinger, J., Floyd, S., Tenenbaum, J. B., & Schulz, L. E. (2017). Children understand that agents maximize expected utilities. *Journal of Experimental Psychology: General*, *146*(11), 1574.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334.

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., Markson, L., & Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PLoS One*, *9*(3), e92160.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299–309.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.* MIT Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT Press.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in Neural Information Processing Systems* (pp. 1874–1882).

Velez-Ginorio, J., Siegel, M., Tenenbaum, J. B., & Jara-Ettinger, J. (2017). Interpreting actions by attributing compositional desires. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society.*

Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. *Science Advances*, *6*(10), eaax5979.

Yildirim, I., Saeed, B., Bennett-Pierre, G., Gerstenberg, T., Tenenbaum, J., & Gweon, H. (2019). Explaining intuitive difficulty judgments by modeling physical effort and risk. *arXiv preprint arXiv:1905.04445*.