Interpreting actions by attributing compositional desires

Joey Velez-Ginorio[1] (joeyv@mit.edu), Max Siegel[1] (maxs@mit.edu), Joshua B. Tenenbaum[1] (jbt@mit.edu), & Julian Jara-Ettinger[2] (julian.jara-ettinger@yale.edu)

[1]Department of Brain and Cognitive Sciences, MIT. Cambridge, MA. 02139
[2]Department of Psychology, Yale University. New Haven, CT. 06520

Abstract

We cannot see others' mental states, so we infer them by watching how people behave. Bayesian inference in a model of rational action – called inverse planning – captures how humans infer desires from observable actions. These models represent desires as simple associations between agents and world states. In this paper we show that by representing desires as probabilistic programs, an inverse planning model can infer complex desires underlying complex behaviors—desires with temporal and logical structure, which can be fulfilled in different ways. Our model, which combines basic desires via logical primitives, is inspired by recent probabilistic grammar-based models of concept learning. Through an experiment where we vary behaviors parametrically, we show that our model predicts with high accuracy how people infer complex desires. Our work sheds light on the representations underlying mental states, and paves the way towards algorithms that can reason about others' minds as we do.

Keywords: social cognition; theory of mind; computational modeling; Bayesian inference.

Introduction

As social creatures, humans routinely have to make sense of what other people are doing, and we do so by appealing to mental states such as beliefs, desires, and intentions. Because we cannot see these internal mental states we need to infer them by watching how people act.

Research into this capacity, called a Theory of Mind (Gopnik & Meltzoff, 1997; Dennett, 1989), suggests that mental state inferences are driven by the assumption that agents act efficiently, subject to constraints imposed by their environment (Gergely & Csibra, 2003). If, for instance, an agent takes a straight path towards a cookie jar, we can guess that her goal is to get a cookie, even before she has reached it. By contrast, if she gets there after wandering around for a while, we may infer that she found it without having deliberately searched.

In such scenarios, it makes sense to equate goals with desires. But in more complex scenarios it is important to distinguish between the two: a one-to-one correspondence between desires and goals is rare. Consider, for instance, if Bob wants to have breakfast. He can do this in several different ways that each require a different plan: he can stay home and prepare breakfast; he could go to the local café near his house; or he could go to a coffee shop that is out of the way. If he chooses to eat at the local café, he can show up and request food. By contrast, if he chooses to cook, he may have to go to the grocery store first and then go to his kitchen, in that order. While Bob is at the grocery store, he may need to buy coffee and milk, but the order in which he buys them does not matter. Finally, before Bob has had breakfast, many states of the world are rewarding (eating at the café or having a scone at home, for example), but once he eats something, all rewards associated with breakfast disappear.

These examples reveal three key properties of desires. First, desires can often be fulfilled in more than one way. So from an observer's standpoint, goals cannot be equated with desires. Second, desires can have logical and temporal structure: they can be fulfilled in different ways (get tea or coffee), they can break into subgoals (get coffee and milk), and they can have temporal structure (go to the café and then buy a scone). Finally, the logical and temporal structure of desires interacts with the underlying rewards. If Bob is thirsty, then both soda and water are rewarding. But once he's had one of them, the other loses its immediate appeal. If Bob wants to exercise and then bathe before work, he has to do them in that specific order; doing them in the wrong order does not suffice. In other cases, the order does not matter, but the reward is only achieved once all the necessary prerequisites are fulfilled. If Bob likes his coffee with milk, then having coffee and milk together is rewarding, but having only one of them is not.

Computational models of mental-state attribution that successfully explain human mental-state inferences assume a relatively simplistic representation of desires: each desire can only be fulfilled in one way, and it is fulfilled by reaching one and only one physical state of the world (e.g. Baker et al., 2017; Baker et al., 2012). This assumption implicitly blurs together desires, intentions, goals, and physical states of the world. As our examples show, this is overly limiting; people may require conjunctions (A and B) or disjunctions of goals (A or B), with temporal properties (A then B).

In this paper we develop a richer representation of desires, and clarify the multiple computational levels that transform desires into actions. To solve the representational challenges, we draw on advances in concept learning that support concepts of unbounded complexity (Piantadosi et al., 2012; Goodman et al., 2008, 2014). To solve the inferential challenges that arise with more sophisticated representations, we draw on advances in mental-state attribution beyond goal inference (Lucas et al., 2014; Jara-Ettinger et al., 2016, *under review*). In the remainder of the paper, we sketch out the computational framework and we present an experiment testing quantitative predictions of our model.

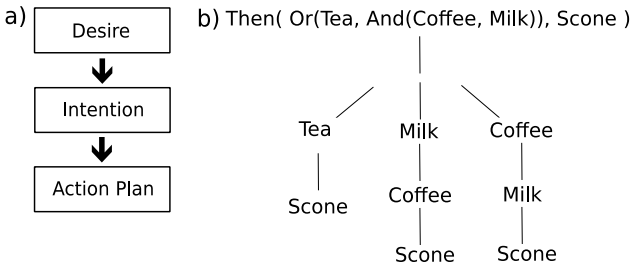

Figure 1. (a) schematic of the generative model. (b) example of how an expression combines primitives and objects to determine how to satisfy a desire. This expression corresponds to an agent who first wants either coffee and milk, or just tea, and then a scone afterwards. The tree below shows the space of possible intentions that can fulfill the desire.

Computational model

We take as a starting point the idea that social cognition is supported by a probabilistic generative model that determines how mental states lead to actions (Baker, et al., 2017). We expand on this approach by building a more powerful representation of desires, and how they relate to behavior.

Figure 1a shows the overall schematic of our model. We argued that a realistic model of commonsense psychology should distinguish between desires, goals, intentions, and actions, and our model attempts to do so.

At the top level we place desires, which combine logical (and/or) and temporal (then) primitives with simple goals (such as arriving to certain physical locations). This approach enables us to represent desires that directly map onto a single goal (e.g. "go to get coffee") as well as desires that can be fulfilled in different ways (e.g. "eat breakfast first, and then either get coffee and milk, or alternatively get tea"). This representation is inspired and based on computational models that combine logical primitives with unitary concepts to explain the productivity and compositionality of conceptual knowledge (Piantadosi et al., 2012; Goodman et al., 2008, 2014).

Following Goodman et al. (2008), we model the space of desires with a probabilistic grammar, which builds arbitrarily complex desires by composing simple ones. The grammar implements production rules that recursively conjoin primitives and units to yield desire expressions. We endow the grammar with several primitives – *And*, *Or*, and *Then* – but the framework is general. These primitives are motivated by common-sense intuitions, but our primary goal is to develop a framework for compositional desires, not to identify the exact primitives that underlie goal-directed behavior.

To connect desires to actions, we rely on an intermediate representation of intentions (see Jara-Ettinger et al., *under review*). Given a composite desire, our model derives the space of intentions as the set of all ordered sequences of sub-goals that satisfy it. For instance, if an agent desires to get either coffee and milk, or just tea, and then a scone afterwards (Fig 1b), her space of intentions is {get tea and then a scone; get milk, coffee, and then a scone; and get coffee, milk, and then a scone}.

To model how the agent selects an intention and transforms it into an action plan, we rely on advances in commonsense psychology that suggest that we interpret other people's behavior through the assumption that they act to maximize their subjective utilities – the difference between the rewards they obtain and the costs they incur (Jara-Ettinger et al., 2016, *under review*; Lucas et al., 2014). This assumption operates at two levels: given a space of intentions, the agent will choose the one that maximizes her subjective utilities, and given an intention, the agent will attempt to complete it as efficiently as possible (for an agent to maximize utilities, they must also minimize costs).

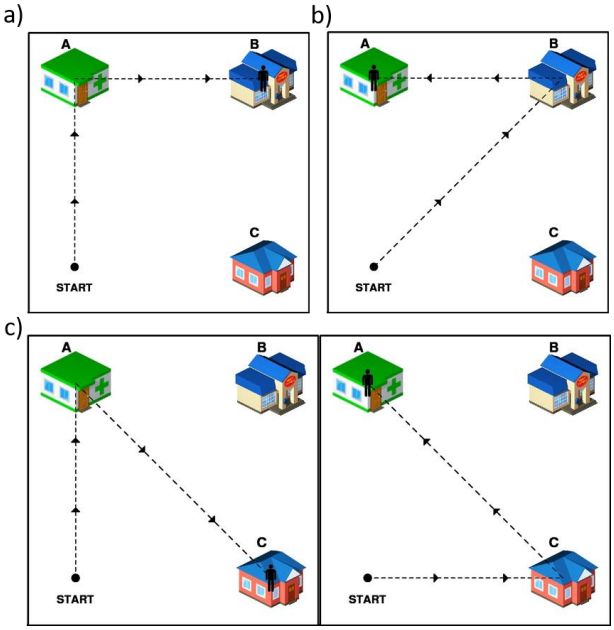

Figure 2: Examples of the experimental stimuli. (a-b) examples of stimuli that consist of a single event. (c) example of stimuli that consists of two events.

To compute each intention's utility, we rely on planning algorithms developed in the robotics literature (Puterman, 2014) that have been successfully applied to model mental-state attribution (Baker et al., 2009; 2017): Markov Decision Processes (MDPs). Given a set of states, a set of actions, and an underlying reward function, MDPs allow us to determine the sequence of actions that an agent should take to fulfill her goal as efficiently as possible. By using MDPs, we can compute the expected cost of achieving each goal, and define an intention's utility as the reward gained by fulfilling the desire minus the sum of the costs for achieving each goal in the intention. Given each intention's utility, we assume that agents probabilistically select an intention:

$$p(I) \propto \exp\left(\frac{U(I)}{\tau}\right) \quad (1)$$

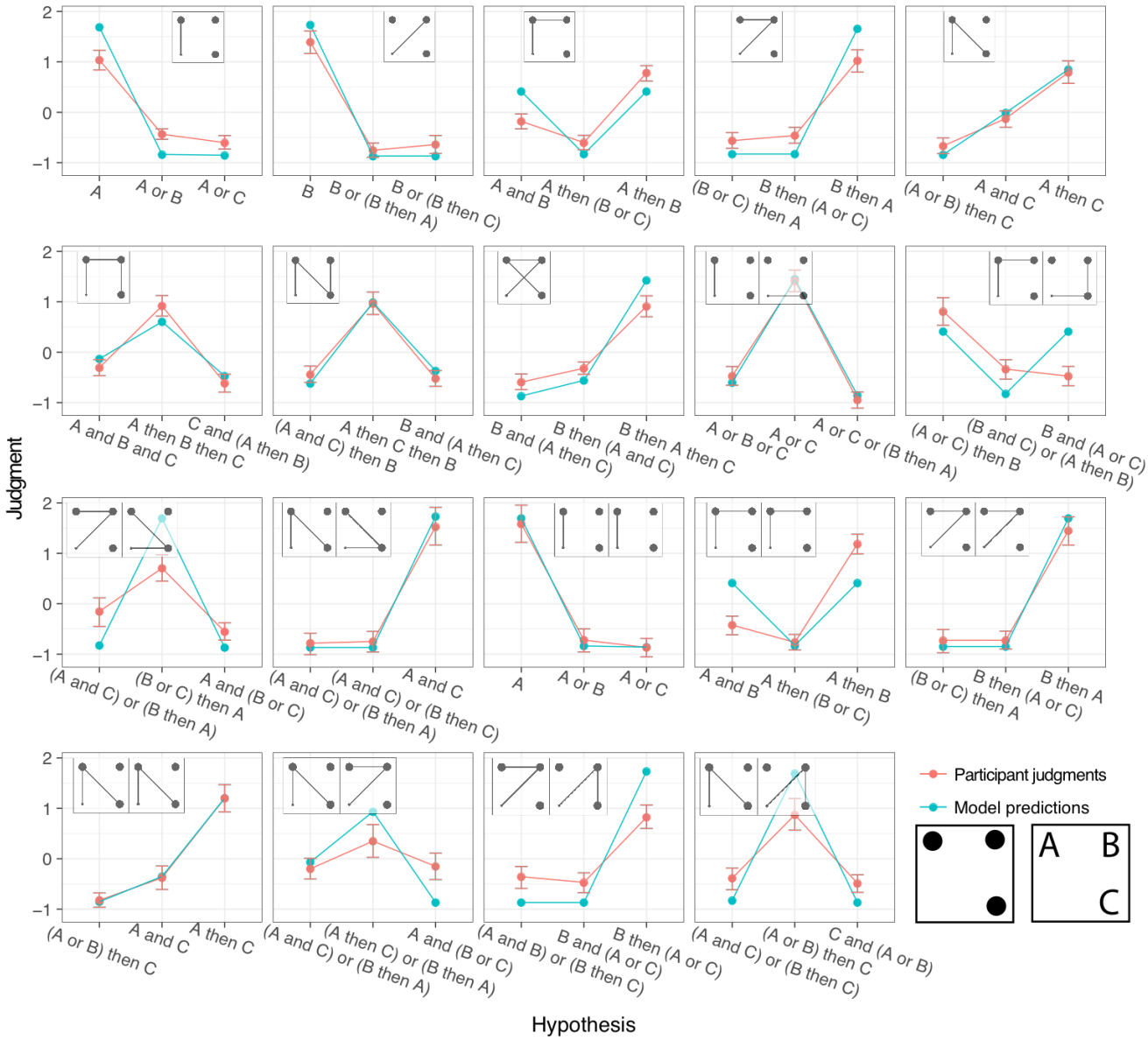

Figure 3: Detailed results from the experiment. Each plot represents one trial from the experiment. The x-axis shows the model's top three hypotheses and the y-axis shows the z-scored prediction with participant judgments. Blue lines and dots show model predictions and red lines and dots show participant judgments. Vertical bars show 95% confidence intervals. In each plot, the schematic represents the paths the agent took in the event (see Figure 2 for examples of the actual stimuli).

where τ is a parameter that captures expectations about the agent's rationality. When τ is low, the agent invariably selects the intention with the highest utility; as τ increases, the agent is more likely to choose a suboptimal intention.

Finally, once the agent has selected an intention, we define the action plan as the ordered sequence of goals along with the motor programs that complete each goal (computed through MDPs).

Inference in the generative model

We have specified a generative model for compositional desires, intentions, and action plans. To recover a desire given some observed actions, we use Bayesian inference to invert the generative model. Given an observable set of actions *A*, the posterior belief for each underlying desire *D* is given by:

$$p(D|A) \propto l(A|D)p(D) \qquad (2)$$

where the prior *p(D)* is set to favor simpler explanations using a simple penalization for the length of the expression (as in Goodman et al., 2008).

To compute the likelihood, *l(A|D)*, we integrate over the space of all possible intentions the agent could have:

$$l(A|D) = \sum_{I \in Intentions} p(A|I)p(I|D) \qquad (3)$$

Both the probability of the intention given the desire ($p(I|D)$), and the probability of the action, given the intention ($p(A|I)$) are computed through the assumption that agents act to maximize their utilities—the difference between the subjective reward for fulfilling their desires minus the cost for fulfilling it. This expectation implies that agents are more likely to act efficiently given their intention, but that they are also more likely to select the intention that can fulfill the desires with the overall lowest cost. We enumerate a set of desires using breadth-first-search over the grammar, and then approximate the posterior over that space using Bayesian inference.

Simplicity prior alternative model

To better understand our model, we developed a simple alternative that uses a deterministic likelihood function, where the probability of a desire generating an action ($p(A|D)$) is 1 if the action satisfies the desire and 0 otherwise. This model continues to have much of the power of the full model: it has access to rich representations of desires and the prior over hypotheses creates a preference for simpler explanations. Unlike the main model, this model is insensitive to the intermediate representations of intentions, as it does not account for how the agent chooses the intention that will fulfill their desires.

Experiment

Design

To evaluate our model, we designed a simple task where participants watched an agent's behavior across one or two days and were asked to determine their belief that the agent had certain desires (see Figure 2).

Methods

Participants 33 participants, mean age (SD) = 32.13 years (9.38 years), range = 20-61 years from the US (as determined by their IP address) were recruited using Amazon's Mechanical Turk Framework.

Stimuli

Figure 2 shows an example of the stimuli. Stimuli consisted of 19 two-dimensional images of an agent traveling to one or more of three potential static locations. Eight of these trials consisted of a single event and the remaining 11 consisted of two events. The one event trials were built by designing all possible efficient paths agents could take to reach between 1 and 3 of the locations and removing equivalent paths (i.e. identical under a rotation or reflection of the map).

Trials with two events were built by first creating a set including possible efficient paths between 1 and 2 of the locations, omitting paths between 3 locations to prevent the stimuli set from growing too large. In contrast to the single event set, we keep the equivalent paths, as they become necessary to construct the most primitive desires occurring over two events e.g. (A *or* B). This creates a base set of 9 paths. To generate the trials with two events, we first split the 9 paths into two classes, one for paths that go to only one location (3 paths) and another for paths that go to two locations (6 paths). For each class we compute the cartesian product of itself, and after removing duplicate pairs of stimuli in each class, (e.g. A,B = B,A), this provided a set of 27 two event trials. From that set, events that violated the principle of rational action were removed (10 trials). Additionally, if a trial with repeated events was the reflection or rotation of another trial with two events, it was removed (5 trials); e.g. between (A,A) and (C,C), we kept (A,A). Last, trials with two events were removed if only one possible hypothesis could explain the trials (2 trials), these trials trials impact our ability to get graded responses on alternative plausible hypotheses (an ideal trial would have more than one plausible explanation, to determine if the model captures the same graded measure humans have for alternatives). For example, if the agent only goes to the farthest location on event 1 and 2, it's clear the only compatible hypothesis is that the agent wants to go that location. As an exception, we included one of these cases in the final set, just to show that the model was capable of inferring the only plausible hypothesis. After filtering the original 27 two event stimuli, 11 remained. These 11 plus the 8 one event trials result in the 19 stimuli used in the experiment.

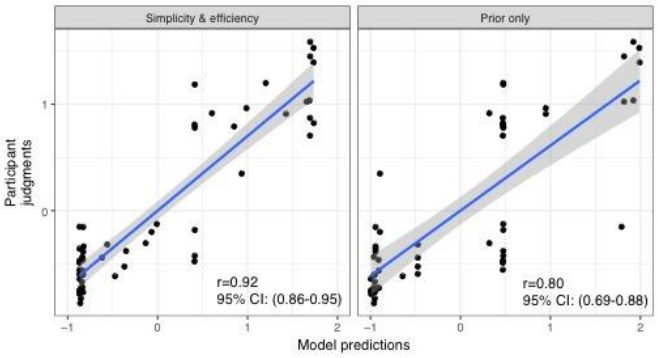

Figure 4. Comparison between our model (simplicity & efficiency) and the alternative model (prior only). Each dot represents a judgment of a hypothesis for a given trial. The x-axis shows the model's prediction and the y-axis shows participant judgments.

Procedure

Participants first read a tutorial that explained the logic of the task. Participants then completed a short survey that ensured they had read the instructions, and the test phase followed immediately after.

During the test phase, participants completed 19 trials. In each trial participants saw the stimuli on the left side, and they were asked to rate their belief that the agent had each of three different desires. Each desire was rated on a scale from 0-10 for each, with 0 indicating "Definitely not"; 5 "Maybe";

and 10 "Definitely." The three desires were obtained by selecting the three hypotheses with the highest posterior distribution according to the model. In order to present these hypotheses to participants, we translated the description from the model into descriptions in English. To ensure their accuracy, two coders blind to the original hypotheses back-translated the descriptions into the model's original representations. The two coders showed 100% agreement and recovered the correct model hypothesis in all trials.

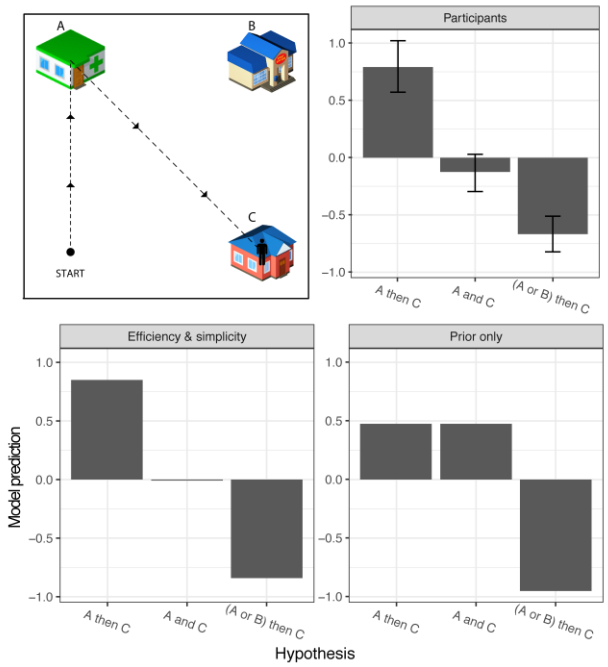

Figure 5. Detailed results one of the trials. The top left plot shows the schematic of the stimuli we used. The top right plot shows participant judgments (z-scored); the bottom two plots show the predictions of the full model and the alternative model (z-scored). This example illustrates how, by removing the probabilistic nature of the likelihood function, the model loses sensitivity to variability in participant judgments.

Results

Figure 3 shows the results from the experiment. Qualitatively, our model fit participant judgments well. Our model predictions showed a correlation of r=0.92 with participant judgments (95% CI: 0.86-0.95). See Figure 4. By contrast, the alternative model (prior only) showed a weaker correlation (r=0.80; 95% CI: 0.69 -0.88). A bootstrap over the correlation difference showed that the full model performed reliably better than the alternative model (correlation difference = 0.11; 95% CI: 0.009-0.18).

Figure 5 shows the detailed results of a single trial that illustrates how the alternative model with a deterministic likelihood function fails to capture participant judgments. In this trial the agent begins by going to the top left location (which is one of the closest ones, together with the bottom right location), and then travels diagonally to the bottom right location. Our full model gives a high probability to the desire that the agent wanted to visit those two locations in that specific order (A *then* C), an average probability to the desire that she could have wanted to visit the locations in any order (A *and* C), and a low probability to the desire that the agent wanted to visit either A or B first, then C ((A *or* B) *then* C). Although all hypotheses explain the actions, our model is sensitive to the probability that each desire would generate the observed actions relative to competing ways to fulfill the same desire (driving the difference between the first and second hypotheses) and to the baseline complexity of the desires (driving the difference between the second and third hypotheses). That is, our model recognizes that there are two equally good intentions that fulfill the desire "A *and* C" (A and then C, or C and then A), but only one that fulfills the ordered desire "A *then* C" (A and then C). This makes our model favor the ordered explanation, as participants do (see Figure 5). This is not captured in the prior only model, as it is only sensitive to expression complexity. These results show how people are both sensitive to the likelihood that a desire would generate the observed actions, and to the complexity of the ascribed desire. Figure 6 shows how this failure becomes even stronger in the case where participants watch the agent behave identically across two events.

Discussion

Here we presented a formal model of action understanding that represents desires as composite entities sampled from a probabilistic context free grammar. Desires get transformed into intentions and then into action plans by the assumption that agents act to maximize their utilities. By performing Bayesian inference over this generative model, we showed how we can capture desires that have rich logical and temporal structure, as well as enabling us to represent desires that can be fulfilled in more than one way. We tested our model by comparing its inferences with those made by human participants, finding that it closely mirrors their judgments, and that an alternative model is less successful.

Our model shows that combinations of primitives and objects using a probabilistic context free grammar supports rich representations of desires in Theory of Mind. The primitives, composing over objects, generate structured desires that capture temporal and logical structure.

Our goal was to develop a more nuanced representation of desires, and the framework we propose works for any arbitrary set of primitives and objects. To test our model, we focused on three specific primitives: *And*, *Or*, and *Then*. Our results do not imply that these are the only primitives people use when they reason about others' desires, or even that they are central in action-understanding. Other primitives such as *If*, *Any*, and *Not*, are likely also at play when we reason about other people's behavior. More research is needed to characterize the primitives we use in action-understanding, and their developmental origins.

To characterize desire complexity, we used a simple prior

that penalized the length of the expression (based on Goodman et al., 2008). Although this is a useful approximation, different primitives may have different priors which capture both their conceptual complexity and the extent to which they are useful in explaining behavior. Future work may attempt to uncover primitive-specific priors and the forces that shape these priors.

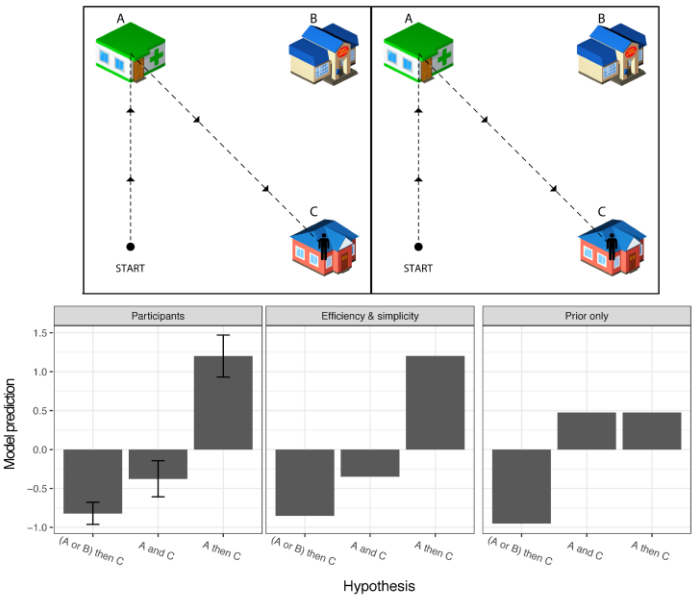

Figure 6. Results from the trial where participants watch two repeated events. While the prior only model continues to make the same predictions, both participants and our model have a stronger belief that the order mattered, in comparison to the trial with a single event (Figure 5)

In our current work, we focused specifically on desires and we assumed that the agents had full knowledge about the environment. In more realistic cases, agents can be uncertain, ignorant, or wrong about the world, and people's reasoning about others is sensitive to this fact (Baker et al., 2017; Kovács, Téglás, & Endress, 2010). Our grammatical approach to desires may also support more structured representations about beliefs. Intuitively, people's beliefs are often structured logically (e.g. my laptop is in my backpack *or* at home; she thinks he is hungry *and* tired). In future work we will investigate the power and limitations of applying this approach to the representations of beliefs, and to the interaction of beliefs and desires.

Although in our work we focused on these representations as applying to desires, these desires often inherit their structure from how the world works. If Bob wants to shoot a water gun, he needs to pour water into the tank first, then pump air into valve, and then press the trigger, in that order. The fact that Bob's desire takes this structure is a reflection of how water guns work. This opens the possibility that, through the ability to reason about other people's desires, we may simultaneously learn procedural knowledge about how to make changes to the world. As such, our model may shed light on how we learn about the world by watching more competent agents (see also Jara-Ettinger, Baker & Tenenbaum, 2012).

Acknowledgments

This work was supported by the Simons Center for the Social Brain. This material is based upon work supported by the Center for Brains, Minds, and Machines (CBMM), funded by NSF-STC award CCF-1231216.

References

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

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 Behavior*

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. Trends in cognitive sciences, 7(7), 287-292.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A Rational Analysis of Rule-Based Concept Learning. *Cognitive Science*, *32*(1), 108-154.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2014). *Concepts in a probabilistic language of thought*. Center for Brains, Minds and Machines (CBMM).

Gopnik, A., Meltzoff, A. N., & Bryant, P. (1997). Words, thoughts, and theories (Vol. 1). Cambridge, MA: Mit Press.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (*under review*). The naïve utility calculus as a rational, quantitative foundation of action understanding.

Jara-Ettinger, J., Baker, C. L., & Tenenbaum, J. B. (2012). Learning What is Where from Social Observations. In *CogSci*.

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.

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*(6012), 1830-1834.

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

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2).

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279-1285.