Learning from Agentic Actions: Modelling Causal Inference from Intention

Dennis W.H. Teo (diswhdt@nus.edu.sg)
Department of Information Systems and Analytics, National University of Singapore

Desmond C. Ong (dco@comp.nus.edu.sg)
Department of Information Systems and Analytics, National University of Singapore
Institute of High Performance Computing, A*STAR Singapore

Abstract

People have the fascinating ability to infer causality by observing other humans' actions. We modelled this process using a Bayesian rational agent model and showed how people can reason about another agent's beliefs and, by extension, infer the world's causal structure. We compared the model's predictions against humans' causal judgements on a novel inference task. Participants (*N* = 171) were shown a dynamic scene depicting either a human agent, robot agent, or both agents acting on two objects sequentially before observing an outcome. After observing the human (vs the less intentional robot) agent, people were more likely to infer that both objects (in sequential order) caused the outcome. When two agents of different intentionality were shown, people favored the object that the intentional agent interacted with as the cause of the outcome. Our model captured these inference patterns well and revealed insights into reasoning about semi-intentional agents and multi-agent contexts.

Keywords: Causal learning; Social inference; Computational cognitive modelling; Bayesian modelling

Introduction

People often learn by observing how other people interact with the world. Take the example of learning to turn on a television set. As a child, we learned how to operate a television remote simply by watching our parents turn on the television. Developmental psychologists argue that this process goes beyond the act of simple imitation (e.g., Lyons, Young, & Keil, 2007; Gardiner, Greif, & Bjorklund, 2011). People actively build causal models of the world and observing the actions of other human beings acts as an important support for this process (Gopnik & Meltzoff, 1998; Waismeyer & Meltzoff, 2017; Meltzoff, Waismeyer, & Gopnik, 2012; Waismeyer, Meltzoff, & Gopnik, 2015). But how does observing other humans actually help learning? How is learning from human interactions different from learning through physical object interactions? In this study, we present a computational model of learning which makes causal inferences from other agents' goal-directed actions and show how this model explains human observers' quick learning of hidden causal relationships.

People can discover hidden causal structures using various methods. Early causal learning research established the importance of spatiotemporal constraints in guiding the learning process (Michotte, 1962; Cohen & Oakes, 1993; Leslie & Keeble, 1987; Oakes & Cohen, 1990). Even without repeated exposure, young children can draw causal inferences between different events as long as they were spatially and temporally close to one another. People can also do probabilistic learning by inferring causality through the natural associations between events (Cheng, 1997; Gopnik, Sobel, Schulz, & Glymour, 2001; Kushnir & Gopnik, 2005). For instance, we can learn about the powers of a television remote by observing how often it is associated with the television turning on. Finally, people also actively take part in the learning process by directly intervening on the world around them (Gopnik & Schulz, 2007; Woodward, 2007, 2005). Rather than passively observing the associations between events, we can explore the buttons of the remote control and see which buttons turn on the television set.

Observational causal learning is a distinct process which allow people to learn by watching other agents interact with the world. Observational causal learning has been shown to be more effective at facilitating causal discovery than probabilistic learning (Meltzoff et al., 2012; Bonawitz et al., 2010) and also does not require the learner to go through (as much) trial and error. Beyond affording learning efficiency, observational causal learning also directs people's attention to learn the most important causal relationships that helps them to navigate their surrounding world (Meltzoff et al., 2012).

Psychologists explain that observational causal learning is guided by intentionality. Similar to how people are more likely to attribute causality to spatially or temporally close events, Meltzoff et al. (2012) proposed that people might also assign greater causal responsibility to intentional actions. Supporting this account, intentional actions has been found to guide the formation of causal beliefs more effectively than either accidental actions (Gardiner et al., 2011; Luchkina, Sommerville, & Sobel, 2018) or spontaneous events (Meltzoff et al., 2012; Bonawitz et al., 2010), especially in contexts of high uncertainty (Gardiner, 2014).

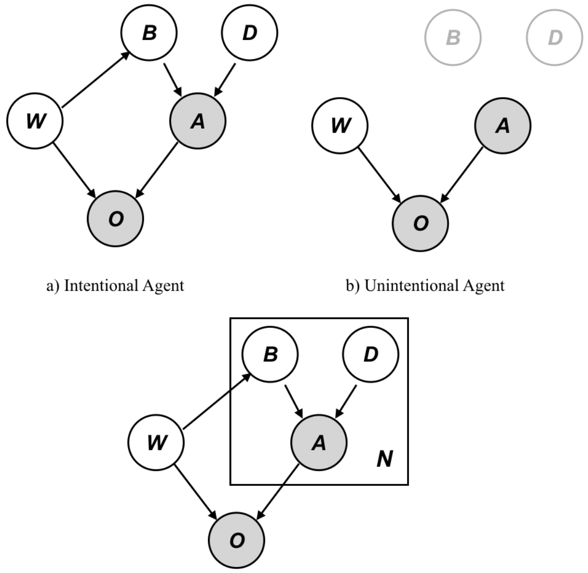
Goodman, Baker, and Tenenbaum (2009) proposed a rational account of how intentionality is used to facilitate causal inference. By assuming that the agent is *goal-directed* and *rational*, the learner can infer the agent's beliefs from their actions which then allows the learner to infer about the external world. For example, when we see our mother push the red button on a remote control, we assume that she is performing that action in order to achieve some outcome, which we soon learn to be turning on the television set. The authors validated their model by comparing its predictions to people's causal inferences of story vignettes describing an actor intentionally intervening upon objects and observing outcomes.

In this paper, we sought to extend Goodman et al.'s (2009) model to better understand how people use intentionality to guide causal inference. In the real world, we often observe entire sequences of causal events (Buchsbaum, Gopnik, Griffiths, & Shafto, 2011) that makes it more challenging to infer the beliefs of the actor and the underlying causal structures. We created a scenario depicting an agent performing a realistic sequence of actions and tasked the model to infer causality from observing the agent's actions. Additionally, we manipulated intentionality by comparing a human agent against a vacuum-bot (although, contrary to our predictions, many participants still perceived the vacuum-bot to be intentional).

We extended Goodman et al.'s (2009) original model in two ways. First, we integrated temporal information and showed how it influences causal inference. When reasoning about an intentional agent, an observer assumes that the agent is actively planning out their actions to achieve their desire. This assumption also takes the temporal order of the planning into consideration. If we observe someone push two buttons (e.g. "A" followed by "B") to turn on the television, we infer that they probably planned to push the two buttons in that sequence. When performing causal inference, we would hence be more likely to infer that "A-then-B" caused the television to turn on rather than "B-then-A". Apart from temporal information from the agent's planning, prior research has also documented causal biases due to temporal contiguity to the observed outcome (e.g., Michotte, 1962; Cohen & Oakes, 1993; Leslie & Keeble, 1987; Oakes & Cohen, 1990). In our model, we integrated both of these processes that make use of temporal information.

Second, we showed how the model can be applied to a multi-agent setting which frequently occurs in the real world. While prior studies have explored how causal inference can be accomplished in a multi-agent context (Maes, Meganck, & Manderick, 2007; Maes, Reumers, & Manderick, 2003), these models were designed to integrate the beliefs and observations of different agents. They do not assume the intentionality of the agents and infer causality based upon the agents' desires, beliefs, and actions. Taking intentionality into account can be especially important in a multi-agent setting where both intentional and unintentional agents are observed. Actions of intentional agents are more informative and are often assigned higher causal weight (Meltzoff et al., 2012). In our model, we propose a computational description of reasoning from multiple agents (intentional and unintentional) within the same context and integrating these inferences to form a coherent causal picture of the world.

We presented the model with a scene depicting an agent (or agents) interacting with two objects in sequence before an automated door opens. The task was to infer the likely cause of the door opening. In the first two conditions (human vs vacuum-bot), the agent acts upon a Blue (B) object followed by a Pink (P) object, before the door opens. If the agents' actions are perceived as intentional, we should expect people to learn that *both* objects B and P (in sequence) caused the door to open. This is because we infer that the agent acted on both B and P *in order* to open the door. However, if the agent is perceived as unintentional, it is less likely that they infer the conjunction of both actions to be the cause. In a third condition, we showed an intentional as well as an unintentional action: the human agent acts on B while the vacuum-bot acts on P. If people are sensitive to intentionality, then we should expect people to assign more causal weight to the human's action than the vacuum-bot's action and infer B to be a more likely cause than P. For each scenario, we compared our model's predictions with the causal judgements of human participants.



a) Intentional Agent

b) Unintentional Agent

c) Two-Agent

Figure 1: Graphical Models. Nodes represent variables, shaded nodes are observable while clear nodes are latent, and edges between nodes represent causal influence. a) Intentional Agent Model, which assumes that outcomes (O) arise from desire-directed (D) actions (A). One can infer the causal structure (W) through agent's beliefs (B). b) Unintentional Agent Model, which assumes that actions are independent of agent's beliefs and desires. c) Two-Agent Model is a mixture of both intentional and unintentional agent models where *N* is the number of observed agents.

Computational Model

We propose a formal model which captures the process of observational causal learning within the context of this study. This model was adapted from Goodman et al. (2009), who integrated a causal Bayesian network with a model of intentional action (e.g., Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). To infer the hidden causal structure of the world, the model makes generative assumptions about how this structure is related to the observed agent's actions as well as any observed changes in the world (See Figure 1a).

In our presented scenario, the model takes the perspective of the observer with the objective of inferring the world's hidden causal structure (W) from the observable actions, A, taken by the agent as well as any observable outcome (O) on the world (i.e. door opening). By assuming that the agent's actions are intentional, the learner assumes that the agent's actions are generated from the agent's beliefs (B) and desires (D). Through these beliefs, the learner can then infer about the likely causal relationships in the world. The process of belief inference can be modelled by the following:

$$P(B|A,D) \propto P(A|B,D)P(B) \quad (1)$$

where we fixed desire (D) to be "open door" since there were no obvious alternative desires in our presented scenario. Then, by assuming that the agent's beliefs are reflective of the underlying world structure ("knowledgeable agent" assumption), the learner can jointly infer the agent's beliefs as well as the causal structures of the world. This is formalized by:

$$P(W,B|A,O,D) \propto P(A|B,D)P(O|W,A)P(B|W)P(W) \quad (2)$$

Causal Structure and Belief, $P(W)$ and $P(B|W)$

We built the hypothesis space using the language of propositional logic (e.g., Goodman, Tenenbaum, Feldman, & Griffiths, 2008). Each hypothesis is a proposition that asserts the necessary actions required to open the door. For example, the hypothesis "interact-with-blue" means that the blue box is necessary to open the door. Each hypothesis can be broken down into atomic actions (e.g. "interact-with-blue") and logical connectors (i.e. *and*, *or*, *then*). To generate the priors for the causal structure, W, we used stochastic recursion to randomly sample and combine the actions with the logical connectives. This can generate conjunctive hypotheses such as "interact-with-blue-*and*-pink". This sampling method allows sampling of arbitrarily complex causal structures, but favors simpler and shorter hypotheses over longer ones. For our model specification, there is a 0.6 probability of sampling an atomic hypothesis, a $0.4 * 0.6$ probability of sampling a first-order conjunctive hypothesis, and a $0.4^2 * (0.6)$ probability of recursion. The beliefs of the agent, B directly inherits the sampled prior of W such that $P(B|W) = P(W)$.

Planning Actions, $P(A|B,D)$

The action space includes a list of actions defined within the setting ("move-to-object-X", "interact-with-object-X", "wait", etc.). $P(A|B,D)$ is estimated by sampling from the action space in accordance to the agent's beliefs (and with desire fixed to "open door"). For instance, the belief "interact-with-blue-and-pink", can generate the action sequence {"interact-with-blue" → "interact-with-pink"} or {"interact-with-pink" → "interact-with-blue"}. Additionally, the sampling is constrained to ensure that the movements are plausible. For instance, an "interact-with-blue" action would be preceded by a "move-to-blue" action.

As people do not always act optimally, we added a number of random actions (from the action space) following an exponential distribution $\varepsilon \sim \lambda e^{-\lambda x}$ at random positions of the action sequence. This reflects the idea that most of the actions are rational (as the mode of the exponential distribution is 0) but humans may, at times, perform unnecessary actions. For our model, we set the free parameter λ to be 1.

Simulating Outcome with Temporal Bias, $P(O|A,W)$

The outcome depends on whether the actions fulfill the requirements of the causal structure. For example, if "interact-with-blue" is the candidate cause, then the door would open if "interact-with-blue" is part of the action sequence.

Additionally, to account for human bias toward events that are temporally closer to each other, we added a decay function which decreases the probability of sampling the outcome as a linear function of the number of "interleaving actions" between the candidate cause and the observed outcome. For example, if the observer samples the causal structure "interact-with-blue-then-pink" and observes the actions {"interact-with-blue" → "interact-with-pink" → "wait" → "move-to-door"} before the door is observed to open, we have two interleaving actions that introduces temporal delay ("wait" and "move-to-door"). This decreases the probability of the door opening by 2δ where δ is the free decay parameter. To choose δ, we performed a grid search using the data from the human agent condition with 5 different values (0, .05, .1, .15, .2) to minimize RMSE. $\delta = .1$ gave the best model fit and this was used to fit the rest of the data. This low δ parameter suggests temporal contiguity bias was weak in the context of this study.

Unintentional Agent Model

When the agent does not appear to be intentional, the model assumes that the observed actions are independent of beliefs and desires (See Figure 1b). This model makes the simple structural assumption that the observed outcomes (O) in the world are determined by the agent's actions (A) and the underlying causal structure (W). The posterior is then just:

$$P(W|O,A) \propto P(O|A,W)P(W)P(A) \quad (3)$$

Two-agent Model

We used a mixture model to model two agents. The model makes independent inferences for each agent and weights the posteriors together. For example, if one intentional agent and one unintentional agent each performs an action, then the model mixes the intentional agent model together with the unintentional agent model to predict the likely causal relationships. The posterior is computed via a weighted sum of the independent predictions of each model.

$$P(W) = \alpha P(W_{intent}) + (1-\alpha)P(W_{unintent}) \quad (4)$$

where α is a free parameter. For this model, we set α to 0.5 to give equal weight to both sets of inferences.

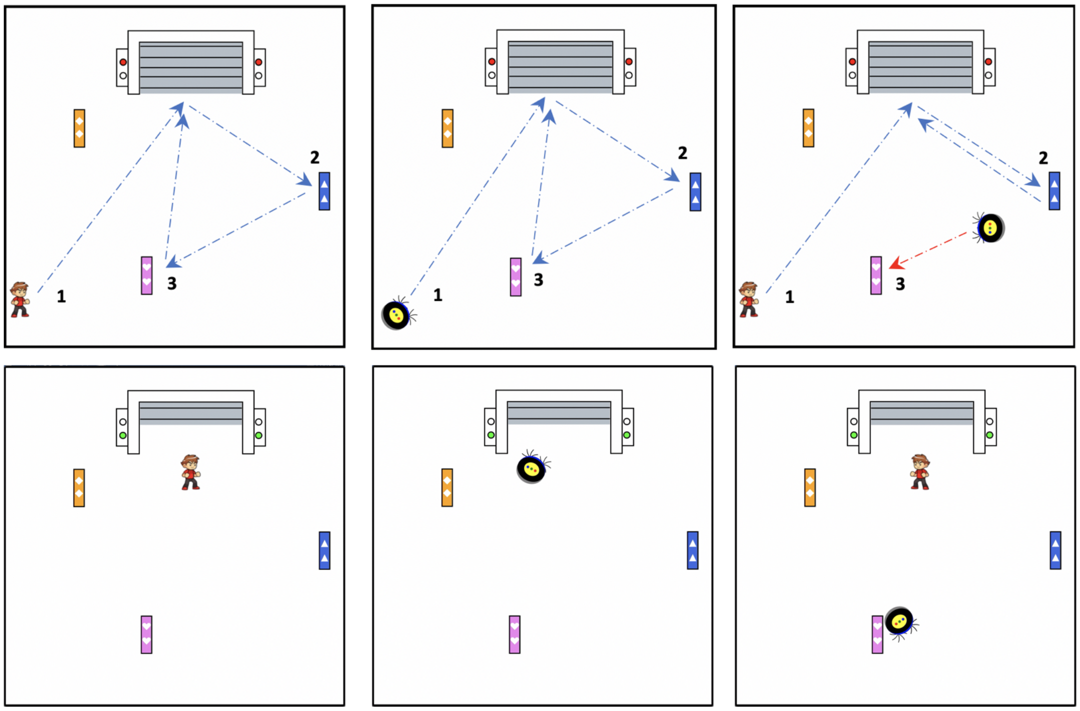


Figure 2: Experimental Materials for the human (left), vacuum-bot (center), and two-agent (right) conditions. Top figures show the movement trajectories of the agents. Agents interact with boxes at positions 2 and 3. The figures below show the positions of the agents right when the door opens. After the door opens, the agents move through the door and the animation ends.

Method

Participants

We collected sample data from 180 participants on Amazon Mechanical Turk, and excluded 9 participants due to flat-line, slow responding, or failing the attention check. The final sample consisted of 171 participants (35% females; 98% native English speakers; $M_{age} = 38.1$; $SD_{age} = 10.0$).

Materials

We constructed 3 video stimuli where an agent moves around a room with 3 boxes (1 blue, 1 pink, and 1 orange) and a metal door (Figure 2). In the human and robot agent conditions, the presented actions were identical, however, we manipulated the agent shown—human vs. vacuum-bot. The agent first moves toward the door before proceeding to push the blue box followed by the pink box. For intentional agents, this action signals their desire to open the door. The boxes make a beeping sound and flash blinking lights after being pushed. After pushing the pink box, the agent then moves toward the door. Nine seconds after the pink box is pushed, the door opens and the agent moves through the door. This time lag exceeds expected delays between contingent events (Shanks, Pearson, & Dickinson, 1989) and ensures that people's causal judgements are less likely biased by temporal proximity.

In the two-agent condition, both a human agent and vacuum-bot were shown. The actions of the human agent were identical to the human condition up to the point where the agent pushes the blue box. After pushing the blue box, he moves straight to the door instead of moving toward the pink box. The vacuum-bot is then shown to push the pink box. Nine seconds after the pink box is pushed, the door opens and the human agent moves through the door.

Procedure

Participants were randomly assigned to the experimental conditions: human agent, robot agent, and two-agent. Depending on the condition, they were shown the corresponding video stimuli described above. After watching the video, participants answered questions based on the video they saw. A single frame of the video was provided below the questions to facilitate recall. Specifically, they were asked how likely they thought that interacting with the different boxes caused the door to open on a Likert scale from 1 (Extremely Unlikely) to 7 (Extremely Likely). They rated a series of different possible combinations (i.e. "blue box only", "pink box only", "blue and pink boxes", "blue then pink box"). They were also asked how likely the agent intended to open the door. For the two-agent condition, this intention attribution question was asked for the human and robot agent separately. Finally, participants were asked for their demographics.

Results

The empirical judgments given by participants and the results from our model are given in the top and bottom rows, respectively, of Figure 3.

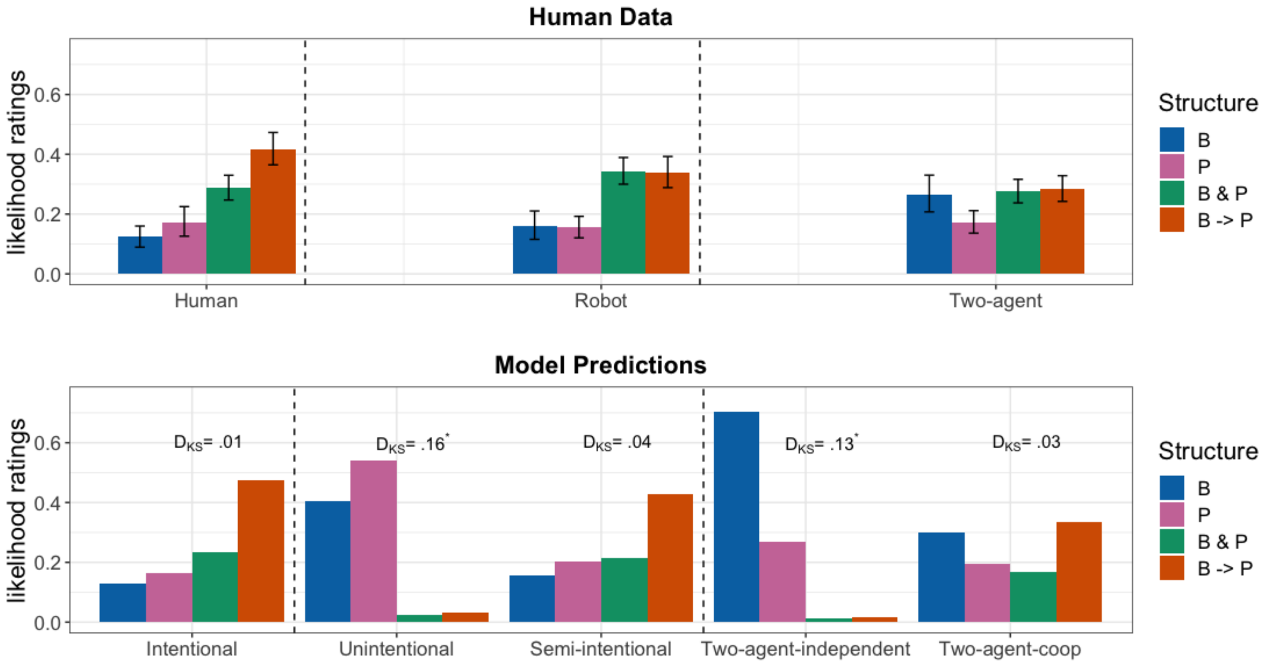


Figure 3: Human Data (top) and Model Predictions (bottom). All ratings were normalised to sum to 1 within each participant. Kolgomorov-Smirnov's D, defined in the text, is a difference measure comparing the model predictions against the human data. The lower the statistic, the more similar the model's predictions are to the human data. K_{DS} ranges from 0 (perfect similarity) to 1 (disjoint distributions). B or P refer to interacting with the Blue or Pink box respectively.

We found that people's causal inferences do, in fact, depend on intentionality. Participants in the human agent condition were more likely to think that "blue-then-pink-box" caused the door to open compared to the robot ($t(110) = 1.99, p = .048$) and two-agent conditions ($t(104) = 2.89, p = .005$). After observing an agent purposefully pushing one box after the other, people were more likely to think that both actions as well as the order of the actions are important. This explanation is further supported by our manipulation check. Comparing the human and robot agent conditions, we confirmed that the vacuum-bot was perceived to be less intentional than the human ($t(96.8) = 3.06, p = .002$). It is worth noting that people still generally regarded the vacuum-bot as intentional (M_{bot} = 5.66 out of 7, SD = 1.59, compared to M_{human} = 6.38 out of 7, SD = .89).

In the two-agent condition, as participants observed the human agent interacting with the blue box, they were more likely to think that the blue box caused the outcome compared to the human ($t(105) = 3.63, p < .001$) and robot agent conditions ($t(116) = 2.63, p = .01$). Again, we found within the two-agent condition that the vacuum-bot was perceived to be less intentional than the human ($t(57) = 5.40, p < .001$).

It is puzzling that participants generally gave the highest likelihood to "both-boxes" and "blue-then-pink-box" across all conditions. In fact, the likelihood ratings for "both-boxes" do not significantly differ across conditions ($F(2,168) = .58, p = .56$). In the next section, we propose a few explanations for these findings and modified our model based on these ideas to compare with the data.

Model Comparison

We implemented the model in WebPPL to obtain model predictions[1]. To assess model fit, we used the Kolgomorov-Smirnov statistic, D_{KS}, as a difference measure between the human data and our model predictions. It is given by the suprenum of the differences (i.e. maximum distance) in cumulative distribution functions of the two distributions:

$$D_{KS} = \sup_{W} |\mathrm{CDF}_{\mathrm{Model}}(W) - \mathrm{CDF}_{\mathrm{Human}}(W)| \tag{5}$$

D_{KS} ranges from 0 to 1; it is 0 when there is complete overlap between the two distributions, and reaches its maximum value of 1 when the two distributions are disjoint.

The intentional agent model predicts that "blue-and-pink-box" and "blue-then-pink-box" are the most likely causes of the door opening as it infers that the agent interacted with the two boxes in order to achieve their goal of opening the door. Comparing our model predictions to the human agent condition (see Figure 3), we see that the intentional agent model fits the data well ($D_{KS} = .01, p = 1$).

[1] Source code is available at `https://tinyurl.com/agenticaction`

However, the unintentional agent model did not fit the participants' ratings well ($D_{KS} = .16, p < .001$). Our unintentional agent model predicts that "pink-box-only" and "blue-box-only" are more likely causes as it tends to favor shorter hypotheses given the lack of an intentional action sequence. The human ratings, on the other hand, give considerably more weight to the conjunction "blue-and-pink" and "blue-then-pink". It is possible that while people see the vacuum-bot as less intentional than the human, they still perceived the vacuum-bot as somewhat intentional. This second explanation is supported by our intention attribution ratings. To model the possibility that people were attributing some intentionality, we used a mixture model of the intentional and unintentional agent models (Fig. 1a and b) to infer causality from the vacuum-bot's actions. This model weights the two possibilities together to infer causality[2]. We found that this model was a better fit of the data ($D_{KS} = .04, p = .99$), which is more evidence that people might have attributed some degree of intentionality to the vacuum-bot.

Finally, we consider the two-agent condition. The two-agent model is itself a mixture of two agents, and makes two *independent* inferences of W given each agent's observations separately. This model predicts that "blue-box-only" is the most likely cause as the intentional agent model assigns high likelihood to the box the human interacted with. Human participants also favored the blue box as a cause over the pink box. However, our model, which we will denote as two-agent-independent in Fig. 3, still did not fit the data well ($D_{KS} = .13, p = .01$). Participants gave high plausibility to the conjunctive hypotheses even when the actions were performed by two separate agents. One possible explanation is that they did not perceive the actions of the two agents to be independent. For example, some people might have thought that pushing the pink box was part of the vacuum-bot's function to aid the human in opening the door. To account for this possibility, we mixed the original two-agent model with another model that assumes the two agents were cooperating with one another[3]. This second model assumes the two agents share joint beliefs (and desire) which account for their combined actions. Structurally, this model is equivalent to the intentional agent model where the belief and actions are treated as if they came from a single agent. This modified two-agent-cooperative model fits the data well ($D_{KS} = .03, p = .99$).

How Intentionality Guides Causal Inference

Our findings show that observational causal learning is well-modelled by a Bayesian rational agent model. When observing an intentional agent, people assume that their actions are not random but determined by the agent's beliefs and planning. This process encodes temporal information and allows people to learn causal chains through observation.

This process may even be used to learn from the actions of non-humans such as robots and animals. Despite our attempts to make the vacuum-bot look unintentional following theories on intentionality perception (Perez-Osorio & Wykowska, 2020), people still attributed high intentionality to the vacuum-bot and inferred causality from its actions similar to that of a human's. This provides evidence that people seem to default toward taking an intentional stance (Dennett, 1989) and readily learn from semi-intentional agents' actions as if they had beliefs and desires.

In the real world, people often observe multiple agents acting upon the world. We showed how humans can still learn from these scenes by performing an inference for each agent and integrating them. During this process, we can also incorporate other assumptions about these agents such as whether they are acting independently or cooperatively (or perhaps even antagonistically; Ullman et al., 2009).

Interestingly, people were inclined to think that the actions of the human and vacuum-bot were *not* independent. One possibility is that people perceived the vacuum-bot to be aiding the human as part of its function. This reference to the *design* of the vacuum-bot is an important element. For example, it is insufficient that observers assume that the human agent has knowledge that the robot agent will push the pink button. This is because this robot agent could have pushed the pink button for any number of reasons and this event would likely not be seen as causally relevant. For the human's and vacuum-bot's actions to be seen as jointly relevant, observers have to assume that the vacuum-bot is pushing the pink button *in order* to help the human agent open the door as part of its design. While this interpretation of the results is intriguing, it is an open question whether it can be generalised. More work needs to be done to see if humans readily ascribe intentional (cooperative) design to other non-human agents.

In this study, we aimed to test the validity and extend the application of Goodman et al.'s (2009) model of observational causal learning. This model is coherent with existing research and was also successful in explaining much of our data. However, it is worth noting that other potential explanations exist. We saw that across all three experimental conditions, people seemed to give high causal weight to the conjunctive hypotheses (e.g. blue-and-pink, blue-then-pink). Instead of arguing for the intentionality of the agents or the dependence between the multiple agents, it might be more parsimonious to assume a human bias toward conjunctive hypotheses. When lacking strong priors of the causal structure, people might view all observed actions as causally relevant similar to the phenomenon of over-imitation (Lyons et al., 2007). Future work should compare these two competing explanations to test if intentionality is really necessary for humans to quickly infer conjunctive causal beliefs.

Conclusion

When we see other agents in the world, we are quick to ascribe intentionality to them. This allows us to better under-

[2] The mixture parameter (α = .9) was chosen by optimizing for the lowest RMSE.

[3] The cooperative model was given a weight of .7 while the original two-(independent)-agent model was given a weight of .3. These free parameters were chosen by optimizing for lowest RMSE.

stand these agents' actions and by extension, to learn about the world. We presented a computational description of how humans make this leap from intention to causality by assuming that these agents are acting rationally and in accordance to their desires. We showed how humans flexibly apply this process to infer causality from semi-intentional agents as well as multi-agent contexts, which has implications for how people learn about their surroundings through social observation.

Acknowledgments

This work was supported in part by a Singapore Ministry of Education Academic Research Fund Tier 1 grant to DCO.

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*(4), 1–10.

Bonawitz, E. B., Ferranti, D., Saxe, R., Gopnik, A., Meltzoff, A. N., Woodward, J., & Schulz, L. E. (2010). Just do it? investigating the gap between prediction and action in toddlers' causal inferences. *Cognition*, *115*(1), 104–117.

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120*(3), 331–340.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367.

Cohen, L. B., & Oakes, L. M. (1993). How infants perceive a simple causal event. *Developmental Psychology*, *29*(3), 421.

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

Gardiner, A. K. (2014). Beyond irrelevant actions: Understanding the role of intentionality in children's imitation of relevant actions. *Journal of Experimental Child Psychology*, *119*, 54–72.

Gardiner, A. K., Greif, M. L., & Bjorklund, D. F. (2011). Guided by intention: Preschoolers' imitation reflects inferences of causation. *Journal of Cognition and Development*, *12*(3), 355–373.

Goodman, N. D., Baker, C. L., & Tenenbaum, J. B. (2009). Cause and intent: Social reasoning in causal learning. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2759–2764).

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.

Gopnik, A., & Meltzoff, A. N. (1998). *Words, thoughts, and theories*. Mit Press.

Gopnik, A., & Schulz, L. (2007). *Causal learning: Psychology, philosophy, and computation*. Oxford University Press.

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, *37*(5), 620.

Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science*, *16*(9), 678–683.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*(3), 265–288.

Luchkina, E., Sommerville, J. A., & Sobel, D. M. (2018). More than just making it go: Toddlers effectively integrate causal efficacy and intentionality in selecting an appropriate causal intervention. *Cognitive Development*, *45*, 48–56.

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences*, *104*(50), 19751–19756.

Maes, S., Meganck, S., & Manderick, B. (2007). Inference in multi-agent causal models. *International Journal of Approximate Reasoning*, *46*(2), 274–299.

Maes, S., Reumers, J., & Manderick, B. (2003). Identifiability of causal effects in a multi-agent causal model. In *Ieee/wic international conference on intelligent agent technology, 2003. iat 2003.* (pp. 605–608).

Meltzoff, A. N., Waismeyer, A., & Gopnik, A. (2012). Learning about causes from people: Observational causal learning in 24-month-old infants. *Developmental Psychology*, *48*(5), 1215.

Michotte, A. (1962). Causalité, permanence et réalité phénoménales.

Oakes, L. M., & Cohen, L. B. (1990). Infant perception of a causal event. *Cognitive Development*, *5*(2), 193–207.

Perez-Osorio, J., & Wykowska, A. (2020). Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, *33*(3), 369–395.

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal contiguity and the judgement of causality by human subjects. *The Quarterly Journal of Experimental Psychology*, *41*(2), 139–159.

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In *Proceedings of the 22nd international conference on neural information processing systems* (pp. 1874–1882).

Waismeyer, A., & Meltzoff, A. N. (2017). Learning to make things happen: Infants' observational learning of social and physical causal events. *Journal of Experimental Child Psychology*, *162*, 58–71.

Waismeyer, A., Meltzoff, A. N., & Gopnik, A. (2015). Causal learning from probabilistic events in 24-month-olds: an action measure. *Developmental Science*, *18*(1), 175–182.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford University Press.

Woodward, J. (2007). Interventionist theories of causation in psychological perspective. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 19–36).