Intuitive Theories

Tobias Gerstenberg (tger@mit.edu) & Joshua B. Tenenbaum

February 29, 2016

Contents

1 Introduction **3**

2 Intuitive theories **4**

2.1 What are intuitive theories? . . . 4

2.2 How can we model intuitive theories? . . . 7

2.3 What are intuitive theories good for? . . . 9

3 Intuitive physics and causal judgments **14**

3.1 Intuitive physics . . . 14

3.1.1 Impetus theory and qualitative reasoning . . . 15

3.1.2 From Noisy Newtons to a mental physics simulation engine . . . 17

3.2 Causal judgments . . . 22

3.2.1 Process vs. dependency accounts of causation . . . 22

Philosophical background . . . 22

Psychological research . . . 23

Bridging process and dependency accounts . . . 27

3.2.2 A counterfactual simulation model of causal judgments . . . 28

Whether-cause . . . 29

How-cause . . . 33

Sufficient-cause . . . 34

Robust-cause . . . 36

3.3 Discussion . . . 36

4 Intuitive psychology and causal explanations **39**

4.1 An intuitive theory of mind . . . 39

4.2 Modeling an intuitive theory of mind . . . 41

4.3 Expressing causal explanations 44

5 Conclusion and future directions **47**

6 Acknowledgments **49**

1 Introduction

Where do babies come from? Why is the sky blue? Why do some people not have enough to eat? Not unlike the most driven scientists, young children have an almost insatiable hunger to figure out how the world works (Frazier, Gelman, & Wellman, 2009). Being bombarded with a series of why-questions by the little ones can be a humbling experience for parents who come to realize their limited understanding of how the world works (Keil, 2003; Mills & Keil, 2004). However, our lack of knowledge about some of the big questions stands in stark contrast to the proficiency and ease with which we navigate our everyday lives. We are remarkably good at filtering out what we really need to know from the vast ocean of facts about the world (Keil, 2012). For example, while most of us are pretty hopeless at explaining how helicopters (or even bicycles) work, we can catch baseballs, pot billiard balls, sink basketballs, or balance a pizza carton on an already overfull trash can hoping that someone else will take out the garbage. Not only can we *do* these things (Todorov, 2004) but also can we make remarkably accurate judgments about these events (see, e.g., Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012), and explain why they happened (Lombrozo, 2012). The Jenga tower fell because you went for the wrong piece. The Red Sox beat the Yankees because their pitcher was tired.

Indeed, the ease with which we sometimes coast through the world can make us blind to the fact that there is something in need of explanation. One way to open our eyes is by learning that some people lack the abilities that we take for granted, such as individuals on the Autism-spectrum who have difficulty understanding the social world (Baron-Cohen, Leslie, & Frith, 1985). Another way is to look at the state of the art in artificial intelligence. In the not too distant future, we will presumably be cruising to work in a self-driving car while experiencing another decisive defeat against the chess application on our phone along the way. These advances are clearly impressive. However, a world like the one portrayed in the movie Her (Jonze (Director), 2013) in which the operating system really *understands* us will most likely remain science fiction for much longer. While Siri, the personal assistant on the iPhone, can tell us where the closest gym is, it cannot answer us who the slacker was in the following sentence: "Tom beat Bill in table tennis because he didn't try hard." (Hartshorne, 2013; Levesque, Davis, & Morgenstern, 2011; Sagi & Rips, 2014) For people, in contrast, the former question may be difficult while the latter is trivially easy – of

course, Bill is the one who didn't try hard rather than Tom.

What explains the huge gap between human and machine intelligence? How can we begin to bridge it? In this chapter, we will argue that understanding common-sense reasoning requires at minimum two key insights: (i) human knowledge is organized in terms of intuitive theories, and (ii) much of human cognition can be understood in terms of causal inferences operating over these intuitive theories. In this chapter, we will focus on two domains of knowledge: people's intuitive understanding of physics and psychology.

The rest of the chapter is organized as follows: We will first clarify what we mean by intuitive theories. We will then discuss how intuitive theories can be modeled within a computational framework and illustrate what intuitive theories are good for. Next, we will put these ideas to work. We will show how people's causal judgments in a physical domain can be explained in terms of counterfactual simulations operating over people's intuitive theory of physics, and how causal explanations of behavior can be understood as inferences over an intuitive theory of mind. We will conclude by discussing some of the key challenges that will need to be addressed to arrive at a more complete understanding of common-sense reasoning.

2 Intuitive theories

2.1 What are intuitive theories?

What do we mean when we say that people's knowledge is represented in the form of intuitive theories? The basic idea is that people possess intuitive theories of different domains, such as physics, psychology, and biology, that are in some way analogous to scientific theories (Carey, 2009; Wellman & Gelman, 1992). Like their scientific counterparts, intuitive theories are comprised of an ontology of concepts, and a system of (causal) laws that govern how the different concepts interrelate. The vocabulary of a theory constitutes a coherent structure that expresses how one part of the theory influences and is influenced by other parts of the theory. A key characteristic of intuitive theories is that they do not simply describe what happened but interpret the evidence through the vocabulary of the theory. A theory's vocabulary is more abstract than the vocabulary that would be necessary to simply describe what happened.

A vivid example for the abstractness of intuitive theories comes from Heider and Simmel's (1944)

seminal study on apparent behavior. Participants who were asked to describe what happened in a movie clip which featured interactions of several geometrical shapes, did not simply describe their movement patterns but rather interpreted the evidence through their intuitive psychological theory and attributed dispositional mental states, such as beliefs and desires to the different shapes. The fact that theories are formulated on a higher level of abstraction allows them to go beyond the particular evidence and make predictions for novel situations. For example, having identified the triangle as mean, allows one to make predictions about how it is likely to behave in other situations. Predictions based on an intuitive theory are intimately linked to explanation (Lombrozo, 2012). Two people who bring different intuitive theories to the same task will reach a different understanding of what happened and make different predictions about the future (maybe the triangle just doesn't like squares but he is generally a nice guy otherwise).

While the concepts and laws in a scientific theory are explicitly defined and known to the scientists in their respective fields, the operation of intuitive theories may be implicit and thus potentially unknown to their user (Borkenau, 1992; Uleman, Adil Saribay, & Gonzalez, 2008). Even though participants in Heider and Simmel's (1944) study described the clips by stipulating specific beliefs and desires, they may not have had complete insight into the workings of their intuitive psychological theory (Malle, 1999). The example further illustrates that our intuitive theories need to be able to cope with uncertainty. A particular action is usually compatible with a multitude of beliefs and desires. If the triangle "runs away" from the square, it might be afraid or it might want to initiate playing catch. Intuitive theories need to embody uncertainty because many inferences are drawn based on limited, and potentially ambiguous evidence.

The example of the moving geometrical shapes also illustrates that intuitive theories postulate latent entities and explain observables (motion patterns) in terms of unobservables (beliefs and desires). An intuitive theory of psychology features observable concepts, such as actions and unobservables, such as mental states. Similarly, an intuitive theory of physics postulates unobservable concepts, such as forces to explain the interaction of observable objects. Importantly, the concepts in an intuitive theory are coherently structured. In the case of an intuitive theory of physics, concepts such as force and momentum are related through abstract laws such as the law of conservation of energy. In the case of an intuitive theory of psychology, beliefs, desires, and actions are linked by a principle of rational action – a person will try to achieve their desires in the most efficient

way possible, given their beliefs about the world (Baker, Saxe, & Tenenbaum, 2009; Dennett, 1987; Wellman, 2011). This principle allows us to make rich inferences based on very sparse data. From observing a person's action (Frank goes to the fridge), we can often infer both their desires and beliefs (Frank is hungry and believes that there is food in the fridge).

Another feature of intuitive theories concerns how they interact with evidence (Gelman & Legare, 2011; Henderson, Goodman, Tenenbaum, & Woodward, 2010). Intuitive theories are characterized by a certain degree of robustness which manifests itself in different ways. Seeing the world through the lenses of an intuitive theory may lead one to simply ignore some aspects that wouldn't be expected based on the theory (Simons, 2000). Further, one's intuitive understanding may lead one to explain away evidence (Nickerson, 1998) or reinterpret what was observed in a way that is theory-consistent (Christensen-Szalanski & Willham, 1991). For example, rather than changing the abstract laws of one's intuitive theory, apparent counterevidence can often be explained by positing unobserved latent causes (Saxe, Tenenbaum, & Carey, 2005; Schulz, Goodman, Tenenbaum, & Jenkins, 2008). When the evidence against one's intuitive theory becomes too strong, one is forced into making sense of the evidence by adopting a new theory (Kuhn, 1996). Some have argued that conceptual changes in development are akin to qualitative paradigm-shifts in science (e.g. Gopnik, 2012).

One of the strongest pieces of evidence for the existence of intuitive theories comes from children's development of a theory of mind. From infancy to preschool, a child's intuitive theory of mind traverses through qualitatively distinct stages. Infants already have expectations about goal-directed actions that are guided by the principle of rational action – an agent is expected to achieve her goals in the most efficient way (Gergely & Csibra, 2003). However, infants form these expectations without yet attributing mental states to agents. Children below the age of four employ an intuitive theory that takes into account an agent's perceptual access and their desires but still fails to consider that an agent's beliefs about the world may be false (Butterfill & Apperly, 2013; Gopnik & Wellman, 1992). Children at this theory of mind stage make systematic errors (Saxe, 2005). Only at around four years of age do children start to realize the importance of beliefs for explaining behavior and that agents can have beliefs that conflict with reality (Perner, Leekam, & Wimmer, 1987; Wellman, Cross, & Watson, 2001).

2.2 How can we model intuitive theories?

We have seen that some of the key properties of intuitive theories are their *abstract structure*, their ability to deal with *uncertainty*, and their intimate relationship with *causal explanation*. How can we best model people's intuitive theories and the inferences they support? What representations and computational processes do we need to postulate in order to capture common-sense reasoning?

In the early days of cognitive science there were two very different traditions of modeling knowledge and inference. Symbolic approaches (e.g. Newell, Shaw, & Simon, 1958) represented knowledge in terms of logical constructs and inference as deduction from premises to conclusions. While these logical representations captured some important structural aspects of knowledge, they did not support inferences in the light of uncertainty. Statistical approaches such as neural networks (e.g. Rumelhart & McClelland, 1988) represented knowledge as statistical connections in the network architecture and inference as changes to these connections. These approaches dealt well with uncertainty but were limited in their capacity to express complex structural relationships. The advent of probabilistic graphical models promised to combine the best of both worlds. Bayesian Networks (BN) integrate structured representations with probabilistic inference (Pearl, 1988). However, none of these approaches was yet capable of representing causality. Pearl (2000) remedied this limitation by developing Causal Bayesian Networks (CBN).

In contrast to BNs where the links between variables merely express statistical dependence, the links in a CBN express autonomous causal mechanisms (Sloman, 2005). Whereas both BNs and CBNs support inferences about unknown variables based on observational evidence about the states of other variables, only CBNs support inferences about what would happen if one were to intervene and change the value of a variable rather than simply observing it. The CBN framework provided a normative account of how people should update their beliefs based on observations versus interventions. Several empirical studies have since established that people are sensitive to this difference (Meder, Gerstenberg, Hagmayer, & Waldmann, 2010; Rottman & Hastie, 2013; Sloman & Hagmayer, 2006; Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005). Inspired by these successes, some have proposed that the CBN framework is a candidate for representing people's intuitive theories (Danks, 2014; Glymour, 2001; Gopnik & Wellman, 2012).

However, a key limitation of CBNs is their limited representational power for expressing ab-

stract, general principles that organize knowledge (Tenenbaum, Griffiths, & Niyogi, 2007). Some of these limitations have been overcome by developing richer frameworks such as hierarchical CBNs that capture causal dependencies on multiple levels of abstraction (Griffiths & Tenenbaum, 2009; Tenenbaum, Griffiths, & Kemp, 2006; Tenenbaum, Kemp, Griffiths, & Goodman, 2011), or frameworks that combine CBNs with first-order logic (Goodman, Tenenbaum, Feldman, & Griffiths, 2008). Even these richer modeling frameworks, however, are insufficient to accommodate two core characteristics of human thought: compositionality and productivity (Fodor, 1975). Like words in language, concepts – the building blocks of thought – can be productively combined in infinitely many ways whereby the meaning of more complex concepts is composed of the meaning of its simpler constituents. We can think and talk about a purple tiger flying through the sky in a small helicopter even though we have never thought that thought before.

Goodman, Tenenbaum, and Gerstenberg (2015) have argued that the compositionality and productivity of human thought can be adequately captured within a framework of probabilistic programs (see also Chater & Oaksford, 2013). Within this framework, a program describes the step-by-step process of how worlds are generated by evaluating a series of functions. The input-output relations between the functions dictate the flow of the program. A function whose output serves as input to another function needs to be evaluated first. What makes a program probabilistic is the fact that randomness is injected into the functions. Thus, each time the program is run, the generative process might take a different route depending on what random choices were made. As a result, the repeated execution of a probabilistic program generates a probability distribution over possible worlds (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008). Gerstenberg and Goodman (2012) have shown how a compact probabilistic program that represents people's intuitive understanding of a simple domain (table tennis tournaments), accurately explains people's inferences based on a multitude of different pieces of evidence (such as how strong different players are based on who beat whom in a series of games).

To make things more concrete, let us illustrate the difference between the representational power of a CBN and a probabilistic program by example of modeling people's intuitive understanding of physics. Sanborn, Mansinghka, and Griffiths (2013) developed a CBN model of how people infer the masses of two colliding objects. Their model incorporated uncertainty about the relevant physical properties and demonstrated a close fit with people's judgments. With a few additions, the model

also captured people's causal judgments about whether a particular collision looked causal or not. The model further explained some deviations of people's judgments from the normative predictions of Newtonian physics – which had traditionally been interpreted as evidence for the operation of heuristic biases – in terms of rational inference on a Newtonian physics model assuming that people are uncertain about the relevant physical properties. These results are impressive. However, at the same time, the scope of the model is quite limited. For example, the model would need to be changed to license inferences about scenes that feature more than two objects. A much more substantial revision would be required if the model were to be used to make predictions about events in two dimensions rather than one. Not only the speed with which objects move and the spatio-temporal aspects of the collision are important for people's causal impression, but also the direction in which the objects are moving after the collision (White, 2012b).

An alternative account for capturing people's intuitive physical reasoning was proposed by Battaglia et al. (2013). They share Sanborn et al.'s (2013) assumption that people's intuitive understanding of physics approximates some aspects of Newtonian mechanics. However, rather than modeling this knowledge in terms of a CBN, Battaglia et al. (2013) stipulate that people's intuitive theory of physics is akin to a physics engine used to render physically realistic scenes. A physics engine is a program designed to efficiently simulate the interaction of physical objects in a way that approximately corresponds to the predictions of Newtonian mechanics. According Battaglia et al.'s (2013) account, people make predictions and inferences about physical events by running mental simulations on their internal physics engine. Assuming that people have uncertainty about some of the relevant parameters then naturally generates a probabilistic program. This approach is not limited to making specific inferences (such as the mass of an object) in specific situations (such as collisions between two objects), but yields predictions about many kinds of questions we might want to ask about physical scenes.

2.3 What are intuitive theories good for?

Now we have a sense of what intuitive theories are and some idea about how they might be modeled in terms of probabilistic, generative programs (for more details, see Goodman et al., 2015). But what are intuitive theories actually good for? Conceiving of intuitive theories in terms of probabilistic, generative models allows to explain a diverse set of cognitive skills as computational

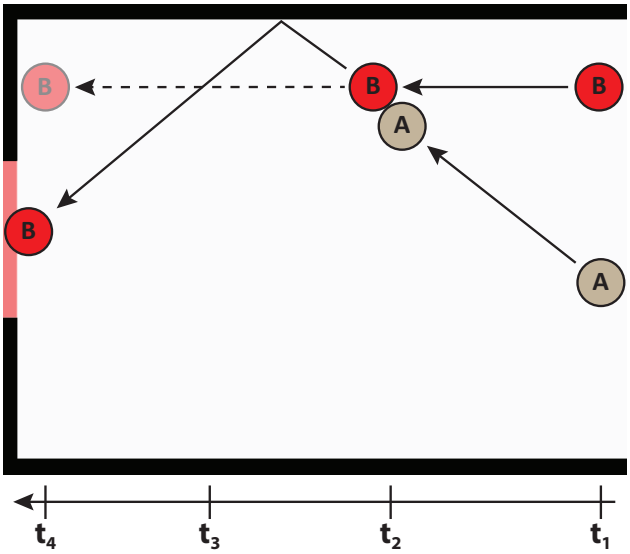


Figure 1: Schematic diagram of a collision event between two billiard balls A and B. The solid lines indicated the balls actual movement paths. The dashed line indicates how ball B would have moved if ball A had not been present in the scene.

operations defined over these programs. Let us illustrate the power of this approach by way of a simple example in the physical domain. Consider the schematic diagram of a collision event between two billiard balls A and B as depicted in Figure 1. Both balls enter the scene from the right at time point t_1. At t_2, the two balls collide. Ball B bounces off the wall shortly afterwards before it eventually enters through an gate in the walls at t_4.

First, intuitive theories support *prediction*. Imagine that the time was stopped at t_2 and we are wondering whether ball B will go through the gate. We can use the generative model to simulate what is likely going to happen in the future by conditioning on what we have observed, such as the state of the table and the trajectories that A and B traveled on up until they collided. Uncertainty enters our predictions in different ways. For example, we might have perceptual uncertainty about where exactly ball A struck ball B. We might also have uncertainty about how ball B is going to collide with the wall. Anyone who has tried a bank shot on a pool table knows that our calculations are sometimes off. However, with practice, accuracy improves dramatically as can be witnessed by watching professional pool players.[1] Finally, we may also have more general uncertainty about

[1]There is also evidence that the way in which novices and experts utilize their intuitive understanding differs. In a recent eye-tracking study (Crespi, Robino, Silva, & de'Sperati, 2012), novices and experts saw video clips of a pool player making a shot. The clip was paused at some point and participants were then asked to judge whether the ball is going to hit a skittle at the center of the table. Novices' eye-movements kept following the path that they predicted the ball will take in an analogous manner. Experts' eyes, in contrast, saccaded quickly from one key spot (e.g. where the ball got struck) to another (e.g. where the ball hits the cushion).

what will happen after this point in time. Will another ball enter the scene and knock ball B off course? Will someone tilt the table? Will the gate suddenly close? All these factors will depend on our more specific understanding of this particular domain, such as whether we've seen other balls entering the scene before or how reliably the gate stays open.

Second, intuitive theories support *inference*. Imagine you checked your phone as the clip started and you only started paying attention at t_2. Having observed the motion paths of the balls from t_2 onwards, allows us to infer where the balls likely came from. Again, we might be somewhat uncertain about the exact location as there are in principle an infinite number of ways in which the balls could have ended up colliding exactly in the way that they did. However, out of all the possible ways, we will deem some more plausible than others (see Smith & Vul, 2014).

Third, we can use our intuitive understanding of the domain as a guide for *action*. Imagine that you are the "gatekeeper" and have to make sure that B doesn't go through. At t_1, you might not see any reason to intervene. However, at t_3 you might get seriously worried that B is actually going to make it this time. Now you might have different actions to prevent that from happening such as throwing another ball, tilting or bouncing the table, or running around the table to catch the ball with your hand. You can use hypothetical reasoning to plan your action so as to minimize effort. If I were to bounce the table from this side, would that be sufficient to divert B so that it won't go through the gate? How hard would I need to bounce the table? Maybe it's safer to throw another ball? But what are the chances that I'm going to actually hit B and knock it off its course?

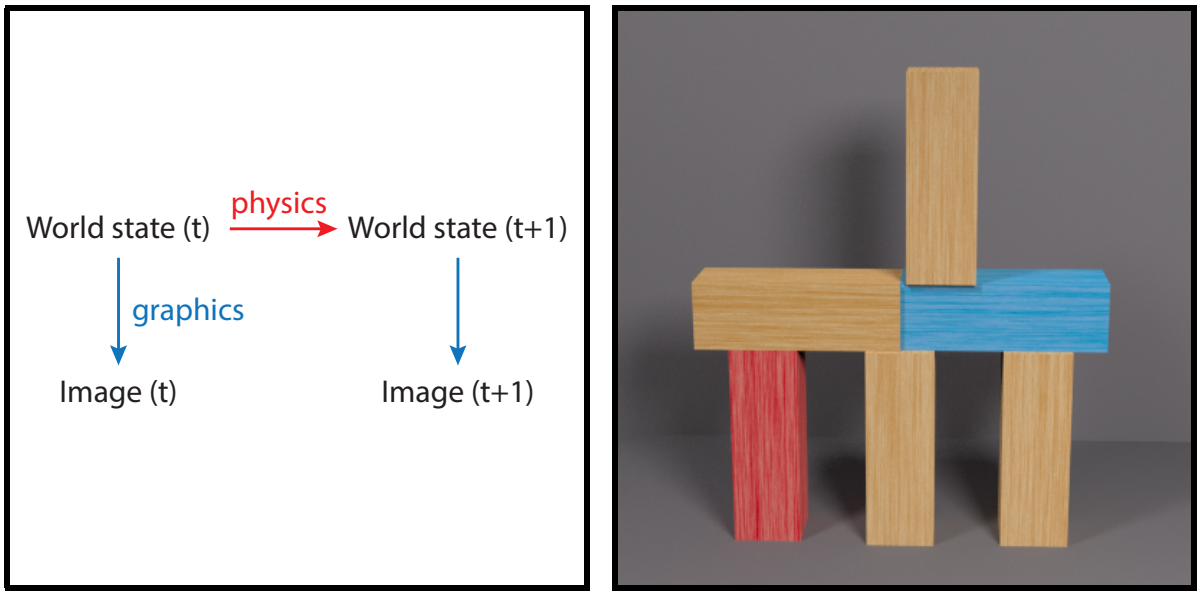
Fourth, generative models support *counterfactual inferences*. For example, having observed what actually happened, we might wonder afterwards what would have happened if the balls hadn't collided. Would ball B have gone through the gate anyhow? Again, we can use our intuitive understanding of the domain to get an answer. We first need to take into account what we've actually observed (as shown by the solid lines in Figure 1). We then realize the truth of the counterfactual antecedent (i.e. that the balls did *not* collide) by means of a hypothetical intervention in the scene. For example, we could imagine that we picked up ball A from the table shortly before the collision would have happened. Finally, we predict what would have happened in this counterfactual scenario. In our case, ball B would have continued on its straight path and missed the gate. Because we have observed the whole episode we can be pretty certain about the counterfactual outcome. We know that no additional balls entered the scene and that noone tilted the table.

Lastly, generative models can be used for *explanation*. If someone asked you why ball B went through the gate, one sensible response would be to say it went through the gate because it collided with ball A. This notion of explanation is tightly linked to causality and, in particular, to a conception of causality which says that what it means to be a cause is to have made a difference in one way or another. In order to figure out whether a particular event made a difference in a given situation, we need to compare what actually happened with the outcome in the relevant counterfactual world. We can cite the collision as a cause for the outcome because it made a difference. If the balls hadn't collided, then ball B wouldn't have gone through the gate. The same operation reveals which things didn't make a difference to the outcome and would thus not satisfy us as explanations of the outcome. For example, imagine someone said that ball B went through the gate because ball A was gray. This explanation strikes us as bad because ball A's color made no difference to the outcome whatsoever. Even if A's color had been yellow instead of gray, ball B would have gone through the gate in exactly the same way.

Explanations not only pick out events that actually made a difference but they tend to pick out "the" cause amongst the multitude of factors that were each "a" cause of the outcome. While it is true that ball B wouldn't have gone through the gate if the top wall hadn't been there, we are less inclined to say that the wall caused the ball to go through the gate (Cheng & Novick, 1992; Hilton & Slugoski, 1986). Explanations distinguish between causes and enabling conditions (Cheng & Novick, 1991; Kuhnmünch & Beller, 2005) and generally pick out events that we consider worth talking about (Hilton, 1990).

In the example that we have used, we employed our generative theory to give an explanation for a particular outcome. We can also use our intuitive theory to provide explanations on a more general level. Imagine that we observed several rounds and noticed that, generally, when both balls are present, ball B tends to miss the gate. In contrast, when ball A is absent B goes through the gate almost all the time. We may thus say, on a general level, that A prevents B from going through the gate. However, even if, A generally tends to prevent B from going through the gate, we would still say that B went through the gate *because* of ball A in the particular example shown in Figure 1. Thus, general causal statements which are based on repeated observations can dissociate from the particular causal statement that pertains to the situation at hand.

So far, we have focused on one particular setup: collision events between billiard balls. However,



(a) Schematic of an intuitive theory of physics.

(b) A tower of blocks.

Figure 2: Using an intuitive theory of physics to understanding a visual scene.

as illustrated above, the power of an intuitive theory that is represented on a sufficiently abstract level (such as an approximate physics simulation engine), is that it supports the same kinds of judgments, actions, and explanations we have described for this particular situation for any kind of situation within its domain of application. There are infinitely many ways in which two billiard balls collide with each other and we ought to be able to say for each situation whether ball A caused ball B to go through the gate or prevented it from going through. And of course, we should also be able to make similar judgments when more than two balls are involved (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015), or other obstacles are present in the scene, such as walls, rough patches, or even teleports (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2014). Indeed, since we have defined the functions operating on intuitive theories simply in terms of conditioning and intervening on a generative model, the same functions can be applied to a completely different context.

Consider the tower of blocks shown in Figure 2b. Using the same intuitive theory of physics (see Figure 2a) we can predict what will happen if the tower is struck from one side (prediction), think about from what direction the wind must have blown after having seeing the top block lying on the ground (inference), put another block on the tower without making it fall (action), consider

what would happen if the red block was removed (hypothetical), and say that the top block doesn't fall because it is supported by the blue block (explanation).

3 Intuitive physics and causal judgments

So far, we have discussed what intuitive theories are, how to model them, and what we can do with them. We have illustrated some of these ideas by thinking about colliding billiard balls and towers of blocks. In Section 3.1 we will use people's understanding of physics as a case study for exploring the ideas behind intuitive theories more thoroughly. We will compare different theoretical approaches to modeling people's intuitive understanding of physics together with some of the empirical studies that have motivated these accounts. In Section 3.2 we will then apply what we have learned about people's intuitive understanding of physics to explain how people make causal judgments.

3.1 Intuitive physics

There is evidence for a foundational understanding of physics from very early in development. Infants already expect that two solid objects cannot occupy the same space and that objects don't just suddenly disappear and reappear but persist over time (Baillargeon, Spelke, & Wasserman, 1985; Spelke, 1990). They infer hidden causes of effects (Saxe et al., 2005) and integrate spatial and temporal information to make predictions about future events (Téglás et al., 2011). Over the course of childhood, our intuitive understanding of physics grows to become more and more sophisticated (for reviews, see Baillargeon, 2004; Spelke, Breinlinger, Macomber, & Jacobson, 1992).

Characterizing people's intuitive understanding of physics is a challenging task (Hayes, 1985). A complete account will have to explain how it is possible for humans to be very apt at interacting with the physical world, while at the same time, when probed more explicitly, some of our intuitive physical concepts appear fundamentally at odds with classical physics (cf. Kozhevnikov & Hegarty, 2001; Levillain & Bonatti, 2011; Shanon, 1976; Zago & Lacquaniti, 2005). In this section, we will first summarize theoretical accounts that have focused on explaining the systematic ways in which people's intuitive physical understanding diverges from the physical laws. We will then discuss more recent work arguing that we can model people's intuitive understanding of physics in analogy to physics engines that are used to create physically realistic animations.

3.1.1 Impetus theory and qualitative reasoning

In the eighties, empirical findings cast doubt on the accuracy of people's intuitive understanding of physics. McCloskey and colleagues revealed several ways in which people's predictions about physical events were off (McCloskey, Washburn, & Felch, 1983; see also, DiSessa, 1982; Zago & Lacquaniti, 2005). In particular, people had difficulty reasoning about projectile motion, such as a when a ball rolls off a cliff (Kaiser, Proffitt, & McCloskey, 1985), or circular motion, such as when a ball whirled at the end of a string is released (McCloskey, Caramazza, & Green, 1980). In the case of projectile motion, many participants tended to draw a path according to which the ball continues its horizontal motion beyond the cliff, and only begins to fall down sometime later. The correct response, however, is that the ball will fall down in a parabolic arc. In the case of circular motion, many participants believed that when the ball is released, it will continue to fly in a curvilinear way before its path eventually straightens out. Here, the correct response is that the ball will fly in a straight line as soon as it's released.

McCloskey (1983) explains people's systematic errors by appealing to a naïve theory of motion. Accordingly, people's intuitive understanding of how objects move is more similar to a medieval impetus theory than to what would be predicted by classical physics. Impetus theory is characterized by two key ideas: first, objects are set in motion by imparting an impetus to the object which subsequently serves as an internal force generating the object's motion. Second, a moving object's impetus gradually dissipates until the object eventually comes to a halt. Impetus theory explains people's answers for the projectile motion and circular motion problems discussed above. By endowing the ball rolling off the cliff with an internal impetus, we can make sense of the ball's initial resistance to gravity. Only when its impetus has dissipated will gravity cause it to fall down. Similarly, if a ball that is whirled has acquired a circular impetus, it takes time for that circular impetus to dissipate before the ball will eventually continue to move along a straight path.

Evidence for people's naïve theory of motion was gathered by having participants predict the motion of objects in diagrams depicting physical scenes, and by having participants explain their responses in extended interview studies. The striking similarities between people's responses in these interviews and the writings of medieval impetus theorists suggests that our naïve theory of motion is likely to be the result of how we experience ourselves as agents interacting with objects

(see also, White, 2012a). For example, people also have the impression that when a moving ball A collides with a stationary B, that A exerted more force on B than vice versa, even though the force transfer is actually symmetrical (White, 2006, 2009). This perceived force asymmetry in collision events may result from people experiencing themselves as agents who exert force on other objects. The experienced resistance by these objects may often be smaller than the experienced force exerted on the object. McCloskey (1983) also argued that people's core theory of motion is surprisingly consistent. Some of the individual differences in people's predictions can be explained as resulting from different beliefs about exactly how impetus dissipates, or how an object's impetus interacts with external forces such as gravity.

Impetus theory draws a qualitative distinction between objects at rest (no impetus), and moving objects that have impetus. In classical physics there is no such distinction. In the absence of external forces, a moving object remains in motion and does not slow down. Fully specifying a physical scene by using the laws of classical physics requires detailed information about the objects and forces at play. However, in many situations, we are able to make qualitative predictions about how a system may change over time without having access to information at a level of detail that would be required to derive predictions based on classical physics. For example, if we put a pot of water on a stove we know that the water will eventually boil – even though we don't know exactly when it will happen. Several accounts have been proposed that aim to capture people's intuitive understanding of physics in terms of qualitative reasoning principles (diSessa, 1993; Forbus, 2010; Kleer & Brown, 1984).

Forbus's (1984) *qualitative process theory* (QPT) states that people's intuitive physical theory is organized around physical processes that bring about qualitatively different states. Accordingly, people's intuitive domain knowledge is represented as a mental model that supports qualitative simulations about the different states a particular physical system may reach. A mental model is characterized by the entities in its domain, the qualitative relationships that hold between the different entities, the processes that bring about change, and the preconditions that must be met for the processes to unfold (Forbus, 1993, 2010; Gentner, 2002). According to QPT, people think about physical systems in terms of qualitative processes that lead the system from one state to another. For example, through increasing temperature water is brought to boil. The differential equations that classical physics requires to model spatio-temporally continuous processes, are replaced with

a qualitative mathematics that yields predictions about how a system may behave based on partial knowledge of the physical scene.

A number of principles guide how people's physical understanding is modeled according to QPT (Forbus, 2010). Rather than representing relevant quantities numerically (such as the amount of water in the pot, or the exact temperature of the stove), qualitative representations are *discretized*. Qualitatively different values are represented that are of *relevance*. For water, its freezing point and boiling point are particularly relevant for understanding its behavior. Qualitative physical models are more *abstract* than their classical counterparts. Instead of precisely quantifying a physical process, qualitative models represent processes in terms of sign changes. For example, when modeling how the water level in a leaking bathtub changes over time when the shower is on, a qualitative model would simply capture whether the level is increasing, decreasing, or constant without representing the exact rate of change. Finally, by abstracting away from more detailed information about the relevant physical variables, a qualitative model often makes *ambiguous predictions*. Qualitative models outline a space of possible states that a system may reach. For the bathtub example, a qualitative model would predict that the bathtub could be completely empty, overflowing with water, or anywhere in between. However, it would not allow us make exact predictions about how the water level would change as a function of the size of the leak and of how much water comes out of the shower.

While classical physical equations such as $F = ma$ are non-causal and symmetric (we could have also written it as $m = \frac{F}{a}$, cf. Mochon & Sloman, 2004), QPT provides an account of people's causal reasoning that is grounded in the notion of a directed physical process that leads from cause to effect. We will discuss process theories of causation in more detail below.

3.1.2 From Noisy Newtons to a mental physics simulation engine

Most of the research reviewed in the previous section has probed people's naïve understanding of physics by asking questions about diagrammatic displays of physical scenes. However, even when dynamic stimuli were used rather than static images, people's judgments in some situations were still more in line with impetus theory rather than what would be predicted by classical physics (Kaiser, Proffitt, Whelan, & Hecht, 1992; Smith, Battaglia, & Vul, 2013).

More recently, research in intuitive physics has revisited the idea that people's understanding

of physics may be best described in terms of some more fine-grained quantitative approximation to aspects of Newtonian physics. Importantly, this research assumes in line with the qualitative reasoning work discussed above, that people have uncertainty about the properties of the physical scene. As briefly mentioned above, Sanborn et al. (2013) have shown how such a noisy Newtonian model adequately captures people's inferences about object masses as well as causal judgments in simple collision events. Their model further explained what was often interpreted to be a biased judgment as a consequence of rational inference over a noisy model that incorporates uncertainty about the relevant physical properties. Michotte (1946/1963) found that people have a stronger causal impression when the velocity of the initially stationary *projectile object* was slightly lower than the velocity of the initially moving *motor object*. Their causal impression was lower when the projectile object's velocity was higher than that of the motor object. Michotte (1946/1963) was puzzled by the fact that people's causal impression wasn't increasing with the magnitude of the effect that the motor object had on the projectile object. However, if we assume that people's intuitive understanding of physics and their causal judgments are closely linked then this effect is to be expected. If both objects are inanimate and on a even surface, it is physically impossible for the projectile object's velocity to be greater than that of the motor object. In contrast, the reverse is possible provided that there is some uncertainty about whether the collision was perfectly elastic. While the collisions of billiard balls are close to being elastic, collisions between most objects aren't and some of the kinematic energy is transformed into heat or object transformation. Thus, the asymmetrical way in which deviations of the projectile object's velocity from the motor object's velocity affect people's causal impressions can be explained as a rational inference in a situation in which we are uncertain about some of the relevant physical properties.

Sanborn et al.'s (2013) Noisy Newton account models people's judgments as inferences over a probabilistic, graphical model that includes variables which express people's uncertainty about some of the parameters. As discussed in the introduction, this model does a very good job of capturing people's inferences about object mass as well as their causal judgments. However, the model is limited in its range of application. Probabilistic graphical models do not generalize well beyond the task that they were built for (cf. Gerstenberg & Goodman, 2012; Goodman et al., 2015; Tenenbaum et al., 2007).

Since then, researchers have explored the idea that people's intuitive understanding of physics

may be best explained in analogy to a physics engine in a computer program that simulates realistic physical interactions. While the Noisy Newton model introduces random variables in the graphical model to capture people's uncertainty, a deterministic physics simulation model can be made probabilistic by introducing noise into the system. For example, when extrapolating a ball's motion, a deterministic physics engine says for each point in time exactly where the ball will be. However, when we try to predict what will happen, we have some uncertainty about exactly where the ball will go (baseball players don't always catch the ball!). By introducing noise into the physics simulation we can capture this uncertainty. Rather than giving an exact value of a where the ball will be at each point in time in the future, a noisy physics simulation model, returns a probability distribution over possible positions. In order to get these probabilities, we generate many samples from our noisy physics engine. Because of the random noise that is injected into the system, each sample looks a little different. The whole sample then induces a probability distribution over possible future states. For example, in the near future, the ball tends to be roughly at the same point in each of our noisy samples since there simply weren't that many steps yet in the simulation to introduce noise. Thus, the noisy simulation model will make a strong prediction about where the ball will be in the near future. However, when asked to make a prediction about where the ball will be later, the model yields a much weaker prediction. Since there were more time steps at which noise was introduced into the system, the outcomes of the simulations are more varied.

Smith and Vul (2013) set out to investigate more closely what sorts of noise in people's mental physical simulations best explains their actions in a simple physics game. In this game, participants saw a moving ball on a table similar to our billiard balls as shown in Figure 1. Part of the table was then occluded and participants were asked to move a paddle up or down such that they will intercept the ball when it reemerges from the occluded part. Smith and Vul (2013) tested different sources of uncertainty: (i) perceptual uncertainty about the ball's position and the direction of its velocity when the occluder occurred, and (ii) dynamic uncertainty about how the ball bounces off the edges of the table and how it moves along the surface of the table. They found that people's actions were best explained by assuming that dynamic noise is a greater factor in people's mental simulations than perceptual noise. For example, the extent to which participants' paddle placement was off increased strongly with the number of bounces that happened behind the occluder and not so strongly with the mere distance traveled. More recently, Smith and Vul (2014) also showed that

the same simulation model also explains people's diagnostic inferences about what path a ball must have taken to arrive at its current position.

In a similar task, Smith, Dechter, Tenenbaum, and Vul (2013) had participants judge whether the ball is going to first hit a green or a red patch on tables with different configurations of obstacles. The earlier participants correctly predicted which patch will be hit, the more reward they received. Hence, participants were encouraged to continuously update their predictions as the clip unfolded. Smith, Dechter, et al. (2013) found that the noisy Newtonian simulation model captured participants' predictions very accurately for most of the trials. However, there was also a number of trials in which it was physically impossible for the ball to get to one of the patches. Whereas participants tended to make their predictions very quickly on these trials, the simulation model took time to realize the impossibility of reaching a certain patch. This result suggests that people sometimes use more qualitative reasoning about what is possible and what is impossible to assist their physical predictions (Forbus, 2010).

While the previous studies focused on people's understanding of collisions in relatively simple 2D worlds, Battaglia et al. (2013) ran a series of experiments to demonstrate different kinds of inferences people make about towers of blocks similar to the one shown in Figure **??**. In one of their experiments, participants saw a configuration of blocks and time was paused. They were asked to judge to what extent they considered the tower to be stable. For trials in which participants received feedback, time was then switched on and participants saw whether the tower was stable or whether some of the blocks fell. Battaglia et al. (2013) modeled judgments by assuming that people have access to an intuitive physics engine (similar to the actual physics engine that was used to generate the simulations) which they can use to mentally simulate what is going to happen. The model assumes that the gradedness in people's judgments stems from perceptual uncertainty about the exact location of the blocks as well as dynamical uncertainty about how exactly the physical interactions are going to unfold.

Battaglia et al.'s (2013) account nicely illustrates some of the key differences between modeling people's intuitive understanding of physics as noisy, mental simulations versus qualitative physical reasoning. The two approaches differ most strongly in the way in which they represent people's uncertainty about the physical scene. The qualitative reasoning approach uses discretization and abstraction to arrive at a symbolic representation that only captures some of the aspects of the

physical situation. In contrast, the noisy simulation approach deals with uncertainty in a very different way. It maintains a richer physical model of the situation and captures people's uncertainty by putting noise on different parameters. In each (mental) simulation of the physical scene, the outcome is determined by the laws of physics as approximately implemented in the physics engine that was used to generate the scene. By running many simulations, we get a probability distribution over possible future scenes because each simulation is somewhat different due to the noise introduced to aspects of the situation that the observer is uncertain about. The more uncertain we are about the physical properties of the scene, the more varied the probability distribution over future outcomes will be.

By expressing uncertainty in this way, the noisy simulation model not only accounts for qualitative judgments such as whether or not the tower is going to fall, but also for judgments that require quantitative precision such as in which direction the tower is most likely to fall. Battaglia et al. (2013) further showed, that neither people nor their model had difficulty doing the same types of judgments when the weights or shapes of the blocks were varied, when physical obstacles where added to the scene, or when they were asked to reason about what would happen if the table that the tower rested on was bumped from different directions. The noisy simulation model also makes predictions on the level of cognitive processing that were recently confirmed. Hamrick, Smith, Griffiths, and Vul (2015) showed that people take longer to make judgments in situations in which the outcome is more uncertain – a finding that fits with the idea that people simulate more (i.e. draw more samples from their mental simulation model) when the outcome is uncertain.

We have now seen some empirical evidence for what an intuitive theory of physics is good for. By assuming that people's intuitive theory of physics is similar to a noisy physics engine, we can explain how people make *predictions* about future events (Battaglia et al., 2013; Sanborn et al., 2013; Smith & Vul, 2012), *inferences* about the past (Smith & Vul, 2014), and take *actions* to achieve their goals (Smith, Battaglia, & Vul, 2013; Smith, Dechter, et al., 2013). In the following we will show that people can also make use of their intuitive theory of physics to reason about counterfactuals and to give explanations for what happened.

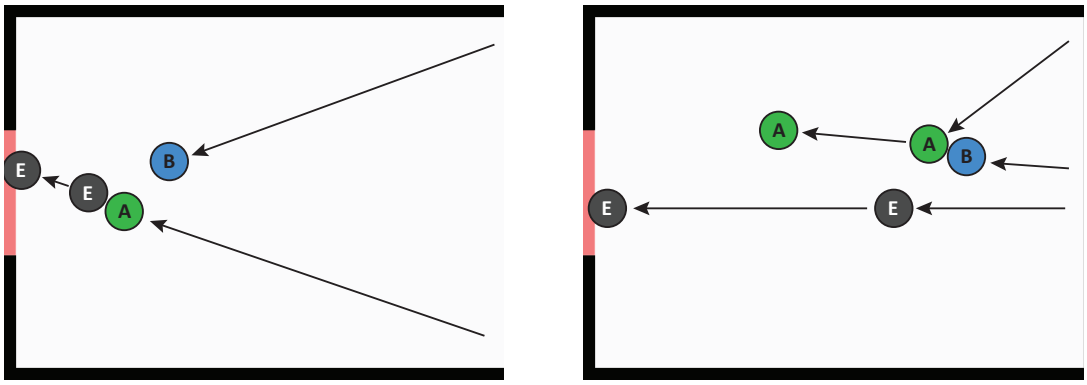
3.2 Causal judgments

We will now shift gears and of focus on applying what we have learned about people's intuitive understanding of physics to explaining how people make causal judgments. Before doing so, let us briefly review some of the philosophical and psychological literature on causality to get a sense for what it is that we need to explain.

3.2.1 Process vs. dependency accounts of causation

Philosophical background In philosophy, there are two broad classes of theories of causation (Beebee, Hitchcock, & Menzies, 2009). According to *process theories* of causation, what it means for an event C to cause another event E is for there to be some physical quantity that is transmitted along a spatio-temporally continuous process from C to E (Dowe, 2000; Salmon, 1984). The paradigm case is a collision of two billiard balls in which ball A transfers its momentum to ball B via the collision event. According to *dependency theories* of causation, what it means for C to cause E is for there to be some kind of dependence between C and E. Some dependency theories propose a probabilistic criterion such that for C to be a cause of E, C must increase the probability that E happens (Suppes, 1970). Here, we will focus on the criterion of counterfactual dependence. According to a counterfactual theory, for C to have caused E both C and E must have happened, and E would not have happened if C had not happened (Lewis, 1973, 1979). The CBN approach we have discussed above is an example of a dependency theory of causation. There, the notion of a counterfactual *intervention* is important: C is a cause of E if E would change in response to an intervention on C (Pearl, 2000; Woodward, 2003). Note that philosophers of causation are not only concerned with providing an account of causation that corresponds to people's intuitive judgments, they care deeply about other aspects as well such as the ontological plausibility of their account, and whether it's possible to reduce causation to counterfactual dependence or certain types of physical processes.

Let us get some intuition about the significance of these different theories by discussing two exemplary cases, each of which is easily dealt with by one theory but is problematic for the other one. Consider the schematic diagram shown in Figure 3a. Both balls A and B enter the scene from the right. Ball E is stationary in front of the gate. Ball A hits ball E and E goes through the gate. Ball B doesn't touch ball E. However, if ball A had been absent from the scene, ball B



(a) Preemption (b) Double prevention

Figure 3: Schematic diagrams of physical interactions between billiard balls.

would have hit E and E would have still gone through the gate, albeit slightly differently. This is a case of *preemption*. The collision event between ball A and E preempts a collision event between B and E that would have happened just a moment later, and which would have resulted in the same outcome. The intuition is that ball A caused ball E to go through the gate whereas ball B did not. Cases of preemption are easily dealt with by process accounts but are problematic for dependency accounts. According to a process theory, A qualifies as the cause because there is a spatio-temporally continuous process through which A transfers momentum to E which results in E going through the gate. In contrast, there is no actual process that connects B and E in any way. Simple dependency theories have now way of distinguishing between the two balls. Ball E would have gone through the gate even if ball A or ball B had been absent from the scene.

Now let's consider the case shown in Figure 3b. Here, all three balls enter the scene from the right. E travels along a straight path and no ball ever interacts with it. However, something interesting happens in the background. Ball A's trajectory is such that it's about to intersect with E. Ball B, however, hits ball A and neither of the balls end up interacting with E. Cases like these are known as situations of *double prevention*. B prevents A which would have prevented E from going through the gate. Clearly, B played an important causal role. Process accounts have difficulty accounting for this since there is no continuous process that connects B and E. Dependency accounts have no trouble with this case. Since E would not have gone through the gate if B hadn't collided with A, B is ruled in as a cause of E's going through the gate.

Psychological research Inspired by the different philosophical attempts of analyzing causality, psychologists have tested which type of theory better explains people's causal learning, reasoning, and attribution. In a typical causal learning experiment, a participant is presented with a number of variables and their task is to figure out what the causal connections between the variables are by observing and actively intervening in the system (Meder et al., 2010; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Waldmann & Hagmayer, 2005). Sometimes participants are already provided with the candidate structure but they are asked to estimate how strong the causal relationships between the variables are (Cheng, 1997; Shanks & Dickinson, 1987; Waldmann & Holyoak, 1992).

The CBN framework provides a unified account for explaining people's judgments about causal strength (Griffiths & Tenenbaum, 2005) as well as their inferences about how different candidate variables are structurally related (for a review, see Rottman & Hastie, 2013). However, there is also evidence that people not only care about the dependency between events when making causal inferences but consider mechanistic information, too (e.g. Ahn & Kalish, 2000; Ahn, Kalish, Medin, & Gelman, 1995). Children, in particular, are more likely to draw conclusions about a causal relationship in the presence of a plausible mechanism (Muentener, Friel, & Schulz, 2012; Schlottmann, 1999). Guided by the normative CBN framework, research into causal learning has mostly focused on providing people with covariation information. However, we know that people use many more sources of information to figure out causal relationships (Lagnado, Waldmann, Hagmayer, & Sloman, 2007). Temporal information is a particularly important cue since causes precede their effects. More recently, research has begun to investigate how people combine the many different sources of evidence to make causal inferences (Bramley, Gerstenberg, & Lagnado, 2014; Lagnado & Sloman, 2004, 2006; Rottman & Keil, 2012).

Besides investigating how people learn about causal relationships, psychologist have also studied how we use our general causal knowledge to make causal judgments about particular events such as in the billiard ball cases shown in Figure 3. Research has shown that people's causal judgments are particularly sensitive to information about causal processes. Several studies have looked into situations in which the predictions of process and dependency accounts are pitted against each other (Chang, 2009; Lombrozo, 2010; Mandel, 2003; Shultz, 1982; Walsh & Sloman, 2011). Based on a comprehensive series of experiments with both adults and children from different cultures,

Shultz (1982) concluded that people's causal judgments are more in line with the predictions of process rather than regularity theories – a particular type of dependency theories. Inspired by early work of Hume (1748/1975), regularity theories predict that people learn about causal relationships by information about covariation and spatio-temporal contiguity. Shultz (1982) found that participants' judgments were more strongly affected by the presence of a plausible mechanism as opposed to other dependency information such as the timing of events.

Similarly, the results of a series of vignette studies by Walsh and Sloman (2011) demonstrated that manipulating process information had a stronger effect on people's cause and prevention judgments than manipulating dependency information. In one scenario, Frank and Sam are playing ball. Frank accidentally kicks the ball toward a neighbor's house. Sam is initially blocking the ball's path but gets distracted and steps out of the way. The ball hits the neighbor's window and smashes it. The majority of participants' (87%) answered the question of whether Frank caused the window to shatter positively, whereas only a small proportion of participants' (24%) agreed that Sam caused the window to smash. While dependency theories have difficulty marking a difference between Frank and Sam (since the outcome counterfactually depended on both of their actions), process theories correctly predict that Frank will seen as a cause but not Sam.

In another series of vignette studies, Lombrozo (2010) found that the actors' intentions had a significant influence on causal judgments in situations of double prevention and preemption. Intentions create a strong dependence relationship between actor and outcome (Heider, 1958; Malle, 2008). If Brian intends to kill Jack and his first shot misses, he is most likely going to shoot another time to achieve his goal. If Brian accidentally shot at Jack and missed, then he certainly won't shoot again. Lombrozo (2010) found that when both the transference cause (equivalent to the (hidden) cause of ball E's motion in our example in Figure 3b) and the dependence cause (equivalent to ball B) were intentional, participants' tended to agree that each of them caused the outcome. However, manipulating intentions had a stronger effect on the dependence cause. While participants' rating of the transference cause was high no matter whether it was intentional or accidental, the dependence cause was seen as less causal when it was accidental rather than intentional.

Research into causal judgments has suffered from a lack of formally specified theories that yield quantitative predictions. Researchers have mostly relied on comparing qualitatively whether causal judgments change between experimental manipulations. Within the class of dependency theories,

the CBN framework has been employed to yield formal definitions of actual causation (Halpern & Pearl, 2005; Hitchcock, 2009) and, more recently, these accounts have been extended to give graded causal judgments by considering default states of variables (Halpern, 2008; Halpern & Hitchcock, forthcoming), or assign degrees of responsibility when multiple causes are at play (Chockler & Halpern, 2004; Lagnado, Gerstenberg, & Zultan, 2013).

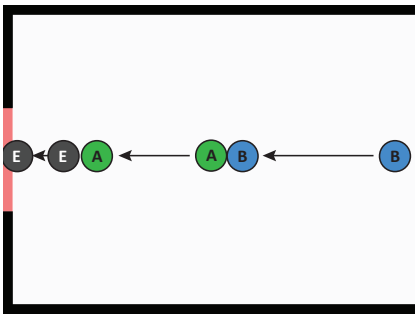
Within the framework of process theories, Wolff (2007) has developed a force dynamics account inspired by work in linguistics (Talmy, 1988). The core idea is that causal events involve the interaction of two parties, an agent and a patient. People's use of different causal terms such as "caused", "prevented", or "helped" is explained in terms of the configuration of forces that characterize the interaction between agent and patient. For example, the force dynamics model predicts that people will say that the agent "caused" the patient to reach an end state, if the patient did not have a tendency toward the end state, the agent and patient forces combined in such a way that the resulting force pointed towards the end state, and the patient actually reached the end state. People are predicted to say "helped" instead of "caused" when the patient already had a tendency toward the end state. In line with Forbus' qualitative reasoning account discussed above, the force dynamic model yields qualitative predictions about what word people should use in a given situation. However, it does not make any quantitative predictions. It is silent, for example, about what makes a really good cause or at what point a "cause" becomes a "helper".

Overall, it is fair to say that existing empirical work on causal judgments doesn't leave us with a very clear picture. Both information about dependence and processes affects people's judgments but the extent to which it does appears to vary between studies. In the case of double prevention, some studies find that participants treat double preventers as causes (Chang, 2009; Lombrozo, 2010; Sloman, Barbey, & Hotaling, 2009; Wolff, Barbey, & Hausknecht, 2010) whereas others don't (Goldvarg & Johnson-Laird, 2001; Walsh & Sloman, 2011). The disparity of the empirical findings reflects the philosophical struggles of finding a unified conception of cause (Paul & Hall, 2013; Strevens, 2013; White, 1990). Indeed, some have given up the hope to find a unified concept of causality and have consequently endorsed the idea that there are two (Hall, 2004) or several fundamentally different concepts of causality (De Vreese, 2006; Godfrey-Smith, 2009; Lombrozo, 2010). Others, in contrast, hold on to the idea that the plurality of causal intuitions can be unified into a singular conception of causality (Schaffer, 2005; Williamson, 2006; Woodward, 2011). In

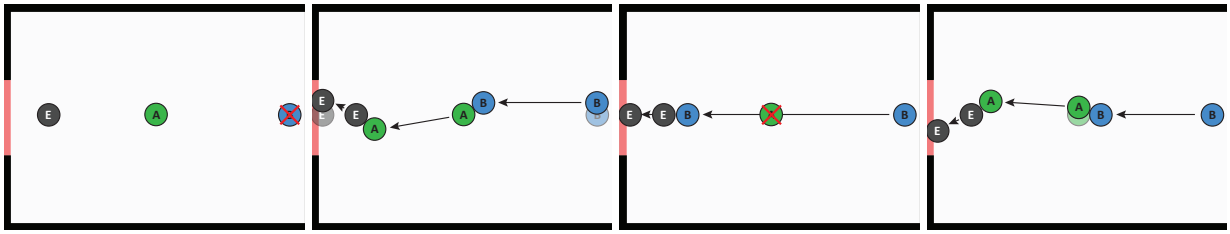
the following, we will argue for the latter position: understanding causal judgments in terms of (different) counterfactual contrasts defined over intuitive theories helps reconcile the different views.

Bridging process and dependency accounts We believe that the notion of causes as difference-makers as conceptualized in dependency theories of causation is primary and that we can capture the intuitions behind process theories of causation in terms of difference-making at the right level of analysis (cf. Schaffer, 2005; Woodward, 2011). Below, we propose an account that is inspired by Lewis' (2000) response to criticisms of his earlier counterfactual theory of causation (Lewis, 1973, 1979). Consider again, the case of preemption as depicted in Figure 3a. It is true that there is no counterfactual dependence between the presence of ball A and whether or not E ends up going through the gate. However, there is a counterfactual dependence on a finer level of granularity – a level that doesn't merely consider absence or presence of the balls but is concerned with the exact way in which the outcome event came about including temporal and spatial information. Lewis (2000) coined this finer notion of counterfactual dependence *causal influence*. Ball A exerts a causal influence on ball E: if ball A had struck ball E slightly differently – at a different angle, with a different speed, or at a different point in time – the relevant outcome event of E going through the gate would have been slightly different, too. E would have gone through the gate at a different location, at a different speed, or at a different point in time. Ball B, in contrast, did not exert any causal influence on ball E on this level of granularity. Even if B's position had been slightly different from what it actually was, E would still have gone through the gate exactly in the same way in which it did in the actual situation.

While we take Lewis' (2000) idea as a point of departure, our proposed account differs from his in two important ways: first, Lewis tried to provide an account that reduces causation to counterfactual dependence and a similarity ordering over possible worlds. In line with more recent work in philosophy of causation (e.g. Woodward, 2003), we believe that causation cannot be reduced but that the concept of actual causation is best understood in terms of counterfactuals defined over an intuitive (causal) theory of the world (Halpern & Pearl, 2005; Pearl, 2000). Second, Lewis believed that conceptualizing causation as influence replaced the earlier idea of thinking about counterfactual dependence on the coarser level of absences and presences. However, we will show that both conceptions of counterfactual dependence are key to understanding people's causal judgments.



(a) actual situation



(b) B=whether-cause? (c) B=how-cause? (d) B=sufficient-cause? (e) B=robust-cause?

Figure 4: Illustration of the different types of counterfactual contrasts that serve as tests to capture different aspects of causation.

3.2.2 A counterfactual simulation model of causal judgments

In the last couple of years, we have developed a counterfactual simulation model (CSM) of causal judgments that aims combine the key insights from process and dependency accounts of causation (Gerstenberg et al., 2012; Gerstenberg, Goodman, et al., 2014, 2015). The CSM starts off with the basic assumption that in in order for a candidate cause (which could be an object or an agent) to have caused an outcome event, it must have made a difference to the outcome. Consider a simple causal chain as shown in Figure 4a. Ball E and ball A are initially at rest. Ball B then enters the scene from the right, hits ball A which subsequently hits ball E, and E goes through the gate. To what extent did balls A and B cause ball E to go through the gate?

Intuitively, both B and A made a difference to the outcome in this situation. The CSM captures this intuition in the following way. For each ball, we consider a counterfactual world in which we had removed the ball from the scene. We then evaluate, using our intuitive physical model of the domain, whether the outcome event would have been any different from what it actually was. More formally, we can express this criterion in the following way:

$$P_{DM}(C, \Delta e) = P(\Delta e' \neq \Delta e | S, remove(C)). \quad (1)$$

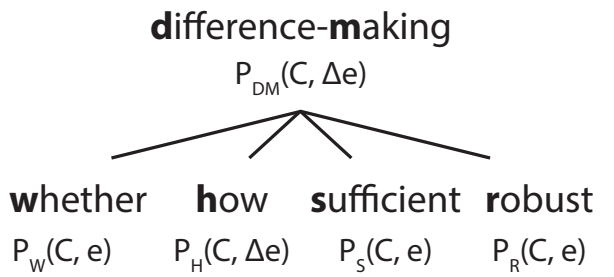


Figure 5: Different aspects of causation that the Counterfactual Simulation Model captures in terms of counterfactual contrasts. Note: The different aspects are defined in Equations 1–5.

To determine our subjective degree of belief that a candidate cause (C) was a difference-maker (DM) for the outcome event (Δe), we first condition on what actually happened in the situation S (i.e. where the balls entered the scene, how they collided, that ball B went through the gate, the position of the walls, etc). We then consider the counterfactual world in which we had removed the candidate cause C from the scene. Then, we evaluate whether the outcome in the counterfactual world ($\Delta e'$) would have been different from the outcome in the actual world (Δe). The Δ sign means that we represent the outcome event of interest on a fine level of granularity. That is, we care about the exact way in which E went through the gate (or missed the gate) which includes spatio-temporal information. It is easy to show that both balls A and B were difference-makers according to this criterion. If ball B had not been present in the scene, then E would not have gone through the gate at all. If we had removed ball A from the scene, E would have gone through the gate differently from how it actually did.

This criterion of difference-making distinguishes candidate objects that were causes of the outcome, from objects that weren't. For example, a ball that is just lying in the corner of the room and never interacts with any of the other balls, would be ruled out. Removing that ball from the scene, would make no difference at all to when and where E went through the gate. If a candidate cause passed this strict criterion of difference-making, then the CSM considers four different aspects of causation that jointly determine the degree to which the candidate is perceived to have caused the outcome of interest (see Figure 5). Let us illustrate these different aspects of causation by focusing on the example of the causal chain.

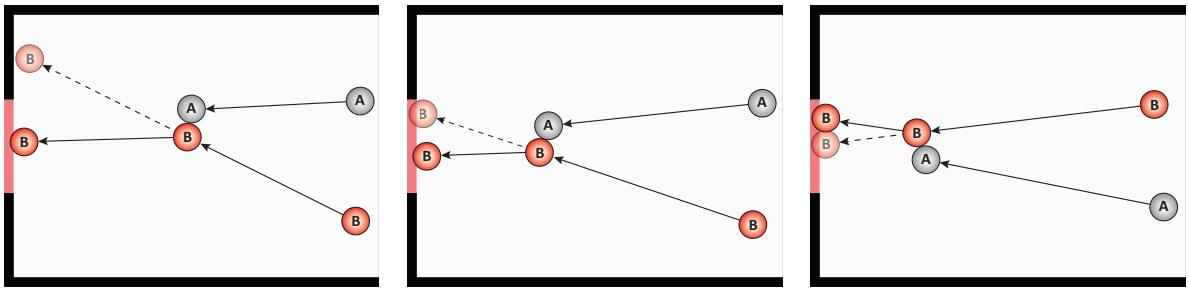
Whether-cause To determine our subjective degree of belief that B was a whether-cause of E's going through the gate (Figure 4b), we consider a counterfactual situation in which B was removed from the scene, and evaluate whether the outcome would have been different from what it actually was:

$$P_W(C, e) = P(e' \neq e | S, remove(C)) \tag{2}$$

Notice that when considering whether-causation, the outcome event (e) is defined at a coarser level of granularity. We are merely interested in whether or not E would have gone through the gate if the candidate cause would have been removed from the scene ($remove(C)$) – we don't care about the more detailed spatio-temporal information. For the causal chain, the answer is pretty simple. E would definitely not have gone through the gate if B had been removed from the scene. Thus, we are certain that B was a whether-cause of E's going through the gate. Ball A, in contrast, was not a whether-cause. E would have gone through the gate even if ball A had not been present.

For the causal chain, determining whether each candidate was a whether-cause of E's going through the gate was easy. However, this need not be the case. Consider the three clips shown in Figure 6. In Figure 6a, it is pretty clear that ball B would have missed the gate if ball A had been removed from the scene. Thus, we are relatively certain that A was a whether-cause of B's going through the gate in this case. In Figure 6b, the situation is less clear. We don't know for sure what the outcome would have been if ball A had been removed from the scene. Finally, in Figure 6c, it is pretty obvious that B would have gone through the gate even if ball A had not been present in the scene. Ball A was not a whether-cause of B's going through the gate in this case.

How can we model people's uncertainty about the outcome in the relevant counterfactual situation in which the candidate cause had been removed from the scene? In line with previous work discussed above, we assume that people's intuitive understanding of this domain can be expressed in terms of a noisy model of Newtonian physics. With this assumption, we can determine the counterfactual probability $P(e' \neq e | S, remove(C))$ in the following way: we generate a number of samples from the physics engine that was used to create the stimuli. Each sample exactly matches what actually happened up until the point at which the two balls collide. At this point, we remove the candidate cause from the scene and let the counterfactual world unfold. For each sample, we



(a) B would have clearly missed.

(b) B would have just missed.

(c) B would have gone in.

Figure 6: Schematic diagrams of collision events. Solid lines show the ball's actual trajectories and the dashed line shows the trajectory ball B would have moved on if it hadn't collided with ball A.

introduce some noise into the underlying physics model by applying a small perturbation to B's direction of motion at each time step. By generating many noisy samples, we get a distribution over the outcome in the counterfactual world. In some of the noisy samples, B goes through the gate, in others it misses. We can then use the proportion of samples in which B ended up going through the gate to predict people's subjective degree of belief that B would have gone through the gate if ball A hadn't been present in the scene.

We have shown that people's causal judgments for clips like the ones shown in Figure 6 are well accounted for by our model of whether-causation (cf. Gerstenberg et al., 2012). In our experiment, participants viewed a number of clips that varied whether B went through the gate or missed it, and how clear it was what would have happened if ball A had not been present in the scene. One group of participants made counterfactual judgments. They judged whether B would have gone through the gate if ball A had not been present in the scene. Another group of participants made causal judgments. They judged to what extent A caused B to go through the gate or prevented it from going through.

We found that participants' counterfactual judgments were very well accounted for by the noisy Newton model. Participants' causal judgments, in turn, followed the predictions of our model of whether-causation very accurately. The more certain participants were that the outcome would have been different if ball A had been removed from the scene, the more they said that A caused B to go through the gate (or prevented it from going through in cases in which it missed). For the clips in Figure 6, participants' causal judgments was high for a), intermediate for b), and low for c).

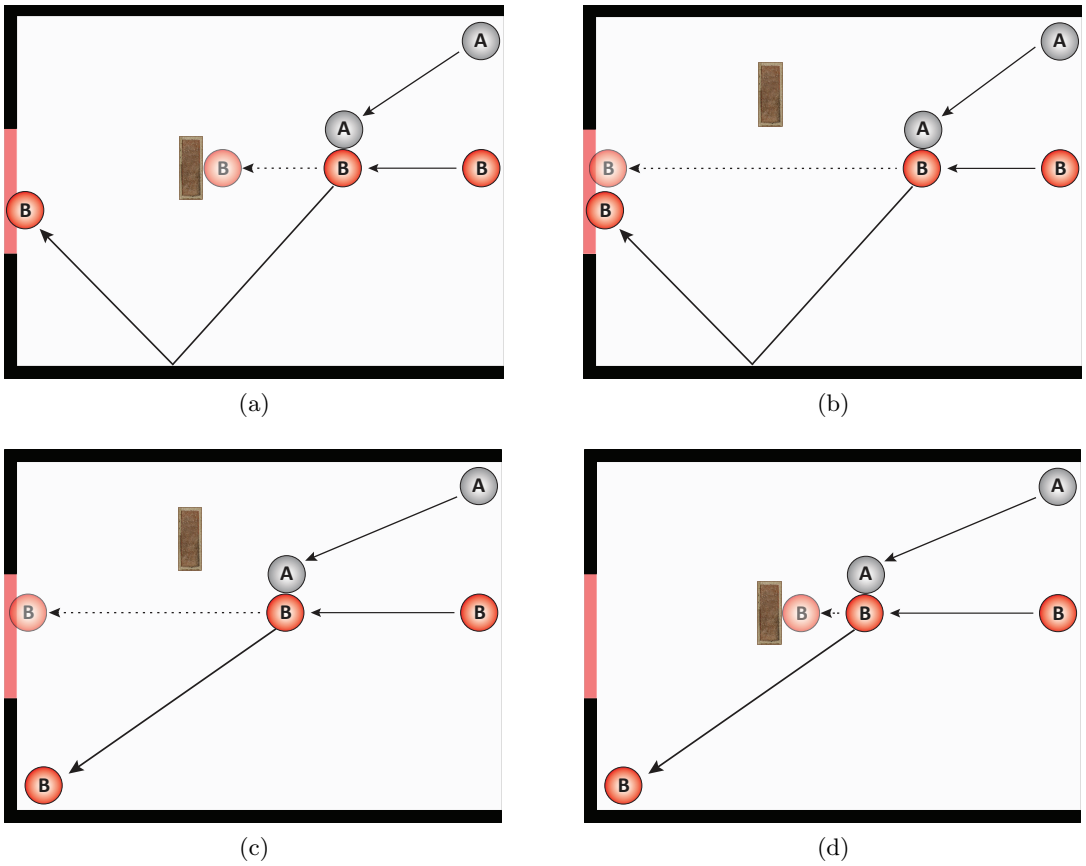


Figure 7: Schematic diagrams of collision events. The balls actual trajectories are shown as solid arrows and the counterfactual trajectory of ball B is shown as dashed arrow. In the top row, B goes through the gate. In the bottom row, B misses the gate. On the left side, A made a difference to the outcome (broadly construed). On the right side, A made no difference.

The tight coupling between causal and counterfactual judgments in Gerstenberg et al. (2012) provides strong evidence for the role of counterfactual thinking in causal judgments. However, since each of the clips that participants saw was somewhat different, it could still be possible, in principle, to provide an account of people's judgments solely in terms of what actually happened.

In another experiment (Gerstenberg, Goodman, et al., 2014), we demonstrated that whether-causation is indeed a necessary aspect of people's causal judgments. This time, we created pairs of clips that were identical in terms of what actually happened, but differed in what would have happened if ball A had been removed from the scene.

Figure 7 shows two pairs of clips. In both clips a) and b), the collision and outcome events are identical. Both clips differ, however, in what would have happened if ball A had not been present in the scene. In a), ball B would have been blocked by the brick. In b), ball B would have

gone through the gate even if ball A had not been present. Participants judged that A caused B to go through the gate for a) but not for b). Similarly, for the two clips shown in c) and d), participants judged that A prevented B from going through the gate in c) but not in d). B would have gone through the gate if A hadn't been present in c). In d), B would have not gone through the gate even if A had not been present – it would have been blocked by the brick. The fact that participants' judgments differ dramatically for clips in which what actually happened was held constant, demonstrates that whether-causation is a crucial aspect of people's causal judgments.

How-cause Some counterfactual theories of causation try to capture people's causal judgments simply in terms of what we have termed whether-causation. Indeed, much of the empirical work discussed above has equated counterfactual theories of causation with a model that merely considers whether-causation, and contrasted this model with process models of causation that are more sensitive to the way in which the outcome actually came about. We believe that the strict dichotomy that is often drawn between counterfactual and process theories of causation is misguided. From the research reported above, it is evident that people care about how events actually came about. However, this does not speak against counterfactual theories of causation. It merely suggests that only considering whether-causation is not sufficient for fully expressing people's causal intuitions. Counterfactual theories are flexible – they can capture difference-making on different levels of analysis. So far, we have focused on whether-causation: the question of whether the presence of the cause made a difference to whether or not the effect of interest occurred. Here, we will show how the CSM captures the fact that people also care about *how* the outcome came about.

Consider again the example of the simple causal chain shown in Figure 4. Participants give a high causal rating to ball B in that case, and an intermediate rating to ball A (see Gerstenberg, Goodman, et al., 2015, for details). If people's causal judgments were solely determined by considering whether-causation, then the fact that A is seen as having caused the outcome to some degree is surprising. The presence of A made no difference as to whether or not E would have gone through the gate. While only B was a whether-cause, both A and B influenced *how* E ended up going through the gate. The CSM defines how-causation in the following way:

$$P_H(C, \Delta e) = P(\Delta e' \neq \Delta e | S, \mathit{change}(C)) \tag{3}$$

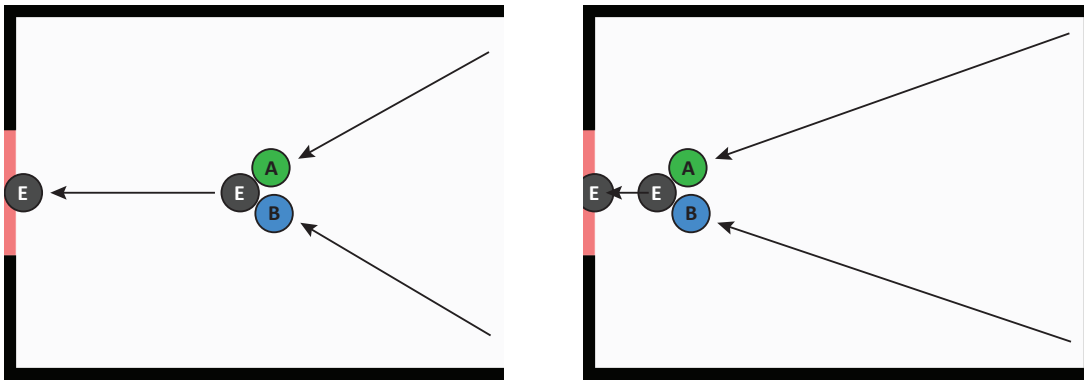
We model our subjective degree of belief that a candidate cause (C) was a how-cause of the outcome, by considering a counterfactual situation in which C was somewhat changed ($change(C)$), and checking whether the outcome (finely construed, that is, including information about when and where it happened) would have been any different in that situation ($\Delta e' \neq \Delta e$).

Notice that in contrast to difference-making and whether-causation, where we considered what would have happened if the candidate cause had been removed from the scene, we need a different kind of counterfactual operation for how-causation. While the *remove* operation is pretty straightforward, the *change* operation is somewhat more flexible. For this particular domain, we can think of the change operation as a small perturbation to the candidate cause's spatial location before the causal event of interest happened (see Figure 4c). A cause is a how-cause if we believe that this small perturbation would have made a difference to exactly how the outcome of interest happened.

By taking into account both whether-causation and how-causation we can make sense of participants' causal judgments in the causal chain. Ball B receives a high judgment because it was both a whether-cause and a how-cause of E's going through the gate. Ball A receives a lower judgment because it was only a how-cause but not a whether-cause of E's going through the gate.

Considering how-causation also allows us to make sense of other empirical phenomena that are troubling for a simple counterfactual model that only relies on whether-causation. While whether-causation and how-causation often go together, they can be dissociated in both ways. In the case of the causal chain, we saw an example where a ball can be a how-cause but not a whether-cause. There are also situations in which a cause is a whether-cause but not a how-cause. In the double prevention clip show in Figure 3b, participants give a relatively low causal rating to ball B even though they are sure that E would not have gone through the gate if ball B had not been present in the scene. In this situation, B was a whether-cause but not a how-cause. Ball E would have gone through the gate exactly in the way in which did even if ball B had been somewhat changed. If people care both about whether-causation and how-causation, then this explains why ball B gets relatively low causal rating in this case.

Sufficient-cause Sufficiency is often discussed alongside necessity as one of the fundamental aspects of causation (e.g. Downing, Sternberg, & Ross, 1985; Hewstone & Jaspars, 1987; Jaspars, Hewstone, & Fincham, 1983; Mackie, 1974; Mandel, 2003; Pearl, 1999; Woodward, 2006). Our



(a) joint causation (b) overdetermination

Figure 8: Schematic diagrams of situations of joint causation (each ball is necessary and both balls are jointly sufficient) and overdetermination (each ball is sufficient).

notion of whether-causation captures the necessity aspect. The CSM defines sufficiency in the following way:

$$P_S(C, e) = P(e' = e | S, remove(\backslash C)) \tag{4}$$

To evaluate whether a candidate cause (C) was sufficient for the outcome (e) to occur in the circumstances (S), we imagine whether the outcome (broadly construed) would still have happened ($e' = e$) even if *all other* candidate causes had been removed from the scene ($remove(\backslash C)$, see Figure 4d). Applying this definition to the causal chain, we see that ball B was sufficient for E's going through the gate. Ball E would have gone through the gate even if ball A had been removed from the scene. Ball A, in contrast, was not sufficient for the outcome. E would not have gone through the gate if ball B had been removed from the scene.

Taking sufficiency into account helps to account for the fact that participants give equally high causal judgments to each ball in situations of joint causation and overdetermination (see Figure 8). A model that only considers how-causation and whether-causation is forced to predict higher causal ratings in the case of joint causation than in the situation of overdetermination. In both situations, the two candidate causes are how-causes of the outcome. In the case of joint causation, both balls are whether-causes but not sufficient-causes of the outcome. In contrast, for overdetermination, both balls are sufficient-causes but not whether-causes of the outcome. If we assume that participants' causal judgments are equally strongly affected by whether-causation and sufficient-causation, we

can make sense of the fact that their judgments are equally high in both cases.

Robust-cause Some causal relationships are more robust than others. Causal relationships are robust to the extent that they would have continued to hold even if the conditions in this particular situation had been somewhat different (cf. Lewis, 1986; Woodward, 2006). The CSM defines robustness in the following way:

$$P_R(C, e) = P(e' = e|S, change(\backslash C)) \tag{5}$$

A candidate cause (C) is a robust cause of the outcome (broadly construed) in the situation (S) to the extent that we believe that the outcome would have still come about ($e' = e$) even if all the other candidate causes had been somewhat different ($change(\backslash C)$). Intuitively, the more factors a particular relationship between a candidate cause the outcome depends on, the more sensitive the cause was.

Taking into account robustness allows us to explain why ball A in the preemption scenario receives such a high rating (Figure 3a). Not only was ball A a how-cause that was sufficient to bring about the outcome. It was also a very robust cause. If we changed the other (preempted) candidate cause, ball A, then ball E would still have gone through the gate exactly in the same way that it did. In contrast in the causal chain scenario (Figure 4a), the initial cause (ball B) was less robust for E's going through the gate. In a counterfactual situation in which ball A had been slightly different, there is a good chance that ball E would not have gone through the gate anymore (Figure 4e).

We have tested the predictions of the CSM in a challenging experiment that included 32 different clips (Gerstenberg, Goodman, et al., 2015). A version of the model that combined how-causation, whether-causation, and sufficiency in a simple additive manner explained participants' causal judgments best. Including robustness as an additional factor in the model, did not improve the model fit significantly.

3.3 Discussion

How do people make causal judgments about physical events? What is the relationship between people's general intuitive understanding of physics and the specific causal judgments they make for

a particular situation?

We have discussed attempts that aim to express people's intuitive understanding of physics qualitatively. More recently, successful attempts have been made to capture people's intuitive understanding of physics as approximately Newtonian. In particular, the assumption that people's intuitive theory of physics can be represented as a probabilistic, generative model has proven very powerful. It explains how people make predictions about the future by sampling from their generative model of the situation, infer what must have happened by conditioning on their observations, and make counterfactual judgments by simulating what the likely outcome would have been if some of the candidate causes had been removed or altered.

We have seen subsequently, how people can use their intuitive understanding of physics to make causal judgments by contrasting what actually happened with the outcome in different counterfactual worlds. Much of previous philosophical and psychological work has argued for multiple notions of causation and explicitly contrasted process theories with dependency theories of causation. We have argued that both views can be reconciled. The counterfactual simulation model (CSM) adequately predicts people's causal judgments for simple collision events by assuming that people's judgments reflect their subjective degree of belief that the candidate cause made a difference to whether the outcome occurred. By contrasting situations in which we matched what actually happened and only varied what would have happened in the relevant counterfactual world, we established that people's causal judgments are intrinsically linked to counterfactual considerations. By looking into more complex scenes which featured several collisions, we showed that people not only care about whether the candidate cause made a difference to whether or not the outcome occurred but also about how the outcome came about and whether the candidate cause was individually sufficient. Good causes are whether-causes, how-causes, and sufficient for bringing about the outcome in a robust way.

The CSM bridges process and dependency accounts in several ways. First, it assumes that people have an intuitive understanding of the physical domain that can be characterized as approximately Newtonian. This generative model specifies the causal laws that are required to simulate what would have happened in the relevant counterfactual world. Second, the CSM acknowledges that people's causal judgments are not simply determined by whether-dependence but are influenced by how-dependence, sufficiency, and robustness as well. Our model is thus closely in line

with a proposal by (Woodward, 2011, p. 409) who argued that "geometrical/mechanical conceptions of causation cannot replace difference-making conceptions in characterizing the behavior of mechanisms, but that some of the intuitions behind the geometrical/mechanical approach can be captured by thinking in terms of spatio-temporally organized difference-making information." In contrast to previous work on causal judgment, the CSM yields quantitative predictions through defining graded concepts of counterfactual contrasts that jointly influence people's causal judgments. The CSM can account for interindividual differences by assuming that people may differ in their assessment of the counterfactual contrasts the model postulates, as well as in how much weight they assign to the different contrasts when judging causation.

The CSM also suggests a new angle for looking at the relationship between language and causation. Recall that Wolff's (2007) force dynamics model explains the use of different causal expressions in terms of differences in force configurations. The difference between "caused" and "helped" is that in the case of "caused" the patient's force was not directed toward the end state whereas in the case of "helped" it was. The CSM suggests different ways in which "helped" (or "enabled") might differ from "caused". First, and similar to the idea in Wolff (2007), people might prefer "helped" to "caused" in situations in which they are unsure about whether the event actually made a difference to the outcome. Indeed, we have shown empirically that if participants believe that the outcome might have happened anyhow even if the causal event hadn't taken place, they prefer to say it "helped" rather than it "caused" the outcome to occur Gerstenberg et al. (2012).

Second, an event might be seen as having "helped" rather than "caused" an outcome, when it was deficient in one way or another. We have seen above that good causes are characterized by whether-dependence, how-dependence, sufficiency, and robustness. For causes for which only some of these factors are true, we might prefer to say "helped" rather than "caused". Consider, for example, the case of double prevention in which ball B is a whether-cause but not a how-cause (see Figure 3b). In this case, people might be more happy to say that ball B "helped" ball E to through the gate gate rather than having caused it to go through. Similarly, ball A in the causal chain is only a how-cause but not a whether-cause. Again, it seems better to say that ball A "helped" ball E to go through the gate rather than "caused" it to go through (cf. Wolff, 2003). The CSM also suggests ways in which "helped" might differ from "enabled".[2] Intuitively, "enabled" is more

[2]Wolff's (2007) force dynamics model does not distinguish between "helped" and "enabled".

strongly tied to whether-dependence. If ball A moves an obstacle out of the way for ball B to go through the gate, it seems appropriate to say that A "enabled" B to go through the gate (A was a whether-cause but not a how-cause). In contrast, if B is already headed toward the gate and A bumps into B to slightly speed it up, it seems like A "helped" rather than "enabled" B to go through the gate (in this case A was a how-cause but not a whether-cause).

4 Intuitive psychology and causal explanations

In the previous sections, we focused on people's intuitive understanding of physics and how it supports people's causal judgments about physical events. We will now turn our attention to people's intuitive understanding of other people.

In an interview with Harvey, Ickes, and Kidd (1978), the psychologist Edward Jones was asked whether the future of attribution theory will see "a convincing integration of cognitive-experimental approaches, such as the Bayesian approach and attributional approaches". Jones's answer was positive: he anticipated an "integration of attribution with information processing, a more mathematical or Bayesian approach." (p. 385). However, this future had to wait. Discouraged by the fate of Bayes' theorem as a seemingly inadequate model of judgment and decision making (e.g. Kahneman, Slovic, & Tversky, 1982; Slovic, Fischhoff, & Lichtenstein, 1977), early Bayesian approaches to attribution theory (Ajzen & Fishbein, 1975, 1983) were met with more critique than approval (Fischhoff, 1976; Fischhoff & Lichtenstein, 1978; Jaspars et al., 1983). Yet, just like Bayesian approaches have had a remarkably successful revival as accounts of judgment and decision making (see, e.g. Hagmayer & Sloman, 2009; Krynski & Tenenbaum, 2007; Sloman & Hagmayer, 2006), they have been rediscovered as powerful accounts for explaining attribution (Hagmayer & Osman, 2012; Sloman, Fernbach, & Ewing, 2012). The anticipated future of an rapprochement between Bayesian and attributional approaches is finally underway.

4.1 An intuitive theory of mind

Heider and Simmel's (1944) experiment in which participants were asked to describe animated clips of moving geometrical shapes is one of the hallmarks of attribution theory. Rather than describing the clip in terms of the shapes' physical movements, most participants explained what had happened by adopting an *intentional* stance (Dennett, 1987; Gergely & Csibra, 2003). As men-

tioned above, most participants perceived the shapes as intentional agents who acted according to their beliefs, desires, and goals. Indeed, many participants reported a rich causally-connected story and endowed the shapes with complex personalities. Developmental psychologists have provided strong empirical support that even infants perceive others as goal-directed agents who are guided by a principle of rational action according to which goals are achieved via the most efficient means available (Gergely, Nádasdy, Csibra, & Bíró, 1995; Scott & Baillargeon, 2013; Sodian, Schoeppner, & Metz, 2004; but see also Ojalehto, Waxman, & Medin, 2013).

How do adults (and infants) arrive at such a rich conception of other agents' behavior? Empirical evidence and theoretical developments suggest that people's inferences about other's behavior are guided both by bottom-up processes, such as visual cues to animacy and intentional action (Barrett, Todd, Miller, & Blythe, 2005; Premack, 1990; Tremoulet & Feldman, 2006; Tremoulet, Feldman, et al., 2000; Zacks, 2004), as well as top-down processes that are dictated by intuitive theories (Uleman et al., 2008; Wellman & Gelman, 1992; Ybarra, 2002). While the fact that top-down processes are required to explain people's inferences has been argued for convincingly (Tenenbaum et al., 2006, 2007, 2011), there is still a heated debate about how these top-down processes feature in people's understanding of other minds (e.g. Stich & Nichols, 1992). According to the *theory theory* (Gopnik & Wellman, 2012, 1992) we understand others by means of an intuitive theory of how mental states, such as desires, beliefs and intentions interact to bring about behavior (cf. Malle, 1999). For example, when we see a person walking towards a hot-dog stand we might reason that she must be hungry and believes that the hot-dog she intends to buy will satiate her desire for food. In contrast, according to the *simulation theory* (Goldman, 2006; Gordon, 1986, 1992) we explain behavior by putting ourselves in the other person's shoes and simulate what mental states we would have if we had acted in this way in the given situation. If I were to walk towards a hot-dog stand I would probably be hungry and intend to get some food. While the last word in this debate is certainly not spoken yet, recent empirical evidence favors the *theory theory* (Saxe, 2005).

Most of the empirical support for the Heiderian view of man (or child) as intuitive theorist comes from developmental psychology (e.g. Gopnik et al., 2004; Gweon & Schulz, 2011; Schulz, 2012; for reviews, see Flavell, 1999; Saxe, Carey, & Kanwisher, 2004). Much of developmental research on theory of mind has focused on the false-belief task (Wimmer & Perner, 1983) in which

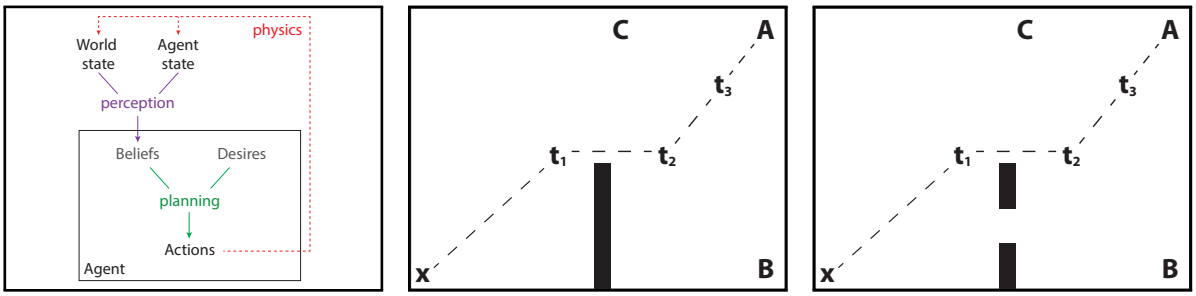
participants are asked to anticipate how an actor will behave whose belief about the state of the world is incorrect (e.g. where Sally will look for a toy which has been moved from one location to another while she was away; see Wellman et al., 2001 for a meta-review). More recently, researchers have begun to also look at the inferences of adult participants in more challenging theory of mind tasks (Apperly, Warren, Andrews, Grant, & Todd, 2011; Birch & Bloom, 2004, 2007; Epley, Keysar, Van Boven, & Gilovich, 2004; Kovács, Téglás, & Endress, 2010).

4.2 Modeling an intuitive theory of mind

A major advantage of the *theory theory* is that it lends itself to a precise computational implementation. In recent years, a number of accounts have been proposed that conceptualize people's inferences about an agent's goals or preferences in terms of an inverse decision-making approach (Baker et al., 2009; Goodman et al., 2006; Lucas, Griffiths, Xu, & Fawcett, 2009; Yoshida, Dolan, & Friston, 2008). Assuming that an intentional agent's actions are caused by their beliefs and desires and guided by a principle of rationality, we can invert this process using Bayes' rule and infer an agent's likely mental states from observing their actions. Building a computational theory of mind is a challenging task because unobservable mental states interact in complex ways to bring about behavior. Any particular action is consistent with a large set of possible beliefs, desires and intentions (see Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015).

These difficulties notwithstanding, Baker et al. (2009) have shown how the inverse planning approach can accurately capture people's inferences about an actor's goals. They use a simplified theory of mind according to which an agent's action is influenced by their beliefs about the environment and their goals. What goals and agent may have is constrained by the setup of the environment. They further make the simplifying assumption that the agent has complete knowledge of the environment (see Figure 9a a schematic of an intuitive theory of agents). The computational task is to infer an agent's goal from their actions in a known environment.

In their experiments, participants observe the movements of an actor in a 2D scene which features three possible goal states (see Figure 9b). Participants are asked to indicate at different time points what they think the agent's goal is. We invite the reader to take a moment before continuing and think about the agent's goals in Figures 9b and 9c at the different time points. In Figure 9b, the agent's goal at t_1 is completely ambiguous. We cannot be sure whether she is



(a) Intuitive theory of agents. (b) Solid barrier (c) Barrier with gap

Figure 9: (a) A simple causal model of the relationships between the *environment*, an agent's *goal* and *action*. While the state of the *environment* and the agent's *actions* are observed the value of the *goal* variable is unknown and needs to be inferred. (b)–(c) Two stimulus examples adapted from Baker et al. (2009). An agent starts at x and moves along the dotted path. Participants are asked about the agent's goal at different time points t_1–t_3.

heading toward A, B, or C. However, at t_2, we can be relatively confident that the agent is not heading toward C. This inference follows from the principle of rationality: if the agent's goal were C then she would have taken a more direct path toward that goal. We are still unsure, however, about whether the agent is heading toward A or B. At t_3 our uncertainty is resolved – as soon as the agent makes another step toward A we are confident that this is the goal she is heading for.

Contrast this pattern of inferences with the situation in which the solid barrier is replaced with a barrier that has a gap (see Figure 9c). In this situation, we rule out B as the agent's goal at t_1 already. If B had been the agent's goal, then she would have walked through the gap in the barrier. This illustrates that our inferences about an actor's goals are not only a function of their actual actions (which are identical in both situations) but markedly influenced by the state of the environment which determines what alternative actions an agent could have performed. In subsequent experiments, Baker et al. (2009) also showed how their account can handle cases in which the agent's trajectory contradicts a simple view of rational action (e.g. when the agent heads toward B at t_2 in Figure 9c) by assuming that the agent's goals might change over time or that the agent might have certain subgoals before reaching the final goal.

The important role that the state of the environment plays in people's attributions resonates well with a core distinction that Heider (1958) drew between what he called *impersonal causation* and *personal causation* (cf. Malle, 2011; Malle, Knobe, O'Laughlin, Pearce, & Nelson, 2000). The key difference between these two notions of causality is the concept of intentional action (see also

Lombrozo, 2010; Woodward, 2006). Whereas an intentional actor adapts to the state of the environment in order to achieve their goal (an instance of *personal causation*) a person who reaches a certain state in the environment *accidentally* would not have reached the same state if the environment had been somewhat different (an instance of *impersonal causation*). While personal causation implies *equifinality* – the same goal is reached via potentially different routes –, impersonal causation (involving physical events or accidental behavior) is characterized by *multifinality* – different environmental conditions lead to different effects.

Further experiments motivated by the inverse planning approach have shown that people are sensitive to configurations of the environment when inferring one agents' social goals of avoiding/approaching (Baker, Goodman, & Tenenbaum, 2008). or helping/hindering another agent (Ullman et al., 2009). While simple social heuristics, such as motion cues, go some way in predicting people's inferences (e.g. avoidance generally motivates an increase in physical distance, cf. Barrett et al., 2005; Zacks, 2004), such accounts are lacking the flexibility to capture the constraints that the environment imposes on behavior. For example, it can sometimes be necessary to walk towards an agent one would like to avoid by fleeing through a door in the middle of a corridor. Furthermore, there are often multiple ways in which one agent can help (or hinder) another agent to achieve their goals (e.g. remove an obstacle, suggest an alternative route, ...). A rational actor will choose the most efficient action in a given situation to realize their (social) goal.

In recent work, Jara-Ettinger, Gweon, Tenenbaum, and Schulz (2015) have extended the inverse planning approach and developed a framework they have coined the *naïve utility calculus*. In addition to inferring an agent's mental states from their actions, we also make inferences about the costs associated with the action, as well as how rewarding the outcome must have been. In a series of experiments, Jara-Ettinger, Gweon, et al. (2015) have shown that children's inferences about the preferences of an agent are sensitive to considerations of agent-specific costs and rewards.

In one of their experiments, an agent chooses between a melon and a banana. On the first trial, the banana is more difficult to get to than the melon because it is placed on a higher pedestal. The agent chooses melon. On the second trial, the difficulty of getting to each fruit is matched. This time, the agent chooses the banana. When five to six year old children are subsequently asked which fruit the agent likes better, they correctly infer that the agent has a preference for the banana. Even though the agent chose each fruit exactly once, children took into account that getting the banana

on the first trial would have been more difficult. The trial in which the costs for both options are equal is more diagnostic for the agent's preference. In another experiment, children made correct inferences about an agents competence based on information about preferences. From observing an agent not taking the preferred treat when it's placed on the high pedestal, we can infer that the agent probably lacks the necessary skill to get it. Jara-Ettinger, Tenenbaum, and Schulz (2015) also showed that children's social evaluations are affected by information about how costly it would be for an agent to help. In situations in which two agents refused to help, children evaluate the less competent agent as nicer. Refusing to help when helping would have been easy reveals more about the person's lack of motivation than when helping would have been difficult.

While most of the previous work assumed that the agent has complete knowledge about the environment, some studies have looked into situations in which the agent can only see a part of their environment. For example, Baker, Saxe, and Tenenbaum (2011) have shown that participants have no difficulty in simultaneously inferring the beliefs and desires of an agent in a partially observable environment. Furthermore, Jara-Ettinger, Baker, and Tenenbaum (2012) demonstrated how from observing other people's actions we cannot only draw inferences about their mental states but also gain useful information about the state of the environment. If we notice how a man gets up from the dinner table next to ours before having finished his meal and walks upstairs, we can use this information to infer the likely location of the bathrooms in the restaurant. How confident we are with our inference will depend on whether or not we think the man has been to the restaurant before (and on whether there might be other reasons for going upstairs such as making an important phone call). More generally, how much we can learn from other's behavior depends on our assumptions about the agent's knowledge state and their intentions (Shafto, Goodman, & Frank, 2012). While assuming that the observed agent has an intention to teach us about the state of the world speeds up learning (Csibra & Gergely, 2009; Goodman, Baker, & Tenenbaum, 2009), we have to remain cautious because intentions can be deceptive (Lagnado, Fenton, & Neil, 2012; Schächtele, Gerstenberg, & Lagnado, 2011).

4.3 Expressing causal explanations

In an insightful epilogue to Jaspars, Fincham, and Hewstone's (1983) volume on attribution research, Harold Kelley argued that the "*common person's understanding of a particular event is*

based on the perceived location of that event within a temporally ordered network of interconnected causes and effects." (p. 333, emphasis in original) Kelley identified five key properties of perceived causal structures that he characterized in terms of the following dichotomies. 1) *simple–complex*: the complexity of the causal relationship between different events encompasses the full range of one-to-one to many-to-many mappings, 2) *proximal–distal*: causes differ in terms of their location on the perceived causal chain of events that connects causes and effects, 3) *past–future*: perceived causal structures are organized according to the temporal order of events and support both reasoning about the past and the future, 4) *stable–unstable*: the causal relationships between events differ in terms of their stability, and 5) *actual–potential*: perceived causal structures not only represent what actually happened but also support the perceiver's imagination about what could have happened.

Thinking of people's intuitive understanding of the physical world and of other agents in terms of intuitive theories resonates very well with Kelley's proposal of perceived causal structures. It highlights that there is no direct mapping between covariation and causal attribution as suggested by early research in attribution theory. Covariation is only one of the many cues that people use in order to construct a causally-structured mental representation of what has happened (Einhorn & Hogarth, 1986; Lagnado, 2011; Lagnado et al., 2007). The perceived causal structure can subsequently be queried, for example by comparing what actually happened with what would have happened under certain counterfactual contingencies, to arrive at causal attributions (Kahneman & Tversky, 1982; Lipe, 1991). Thinking about causal attributions in these terms shifts the focus of interest toward what factors influence people's causal representations of the world such as temporal information (Lagnado & Sloman, 2004, 2006) and domain knowledge (Abelson & Lalljee, 1988; Bowerman, 1978; Kelley, 1972; Mischel, 2004; Schank & Abelson, 1977; Tenenbaum et al., 2007).

We have shown above how the Counterfactual Simulation Model (CSM) adequately captures people's causal judgments about collision events. The different aspects of causation in the CSM are defined on a sufficiently general level such that they can be applied to any generative model of a domain – including people's intuitive theory of psychology (Mitchell, 2006; Wellman & Gelman, 1992). Consider the scenario depicted in Figure 10 (cf. Baker et al., 2011). An agent is about to grab some food for lunch from a food truck. The agent knows that there are three different food trucks: one with Mexican food, one with Lebanese food, and one with Korean food. However, there are only two parking spots which are taken on a first-come, first-served basis. Baker et

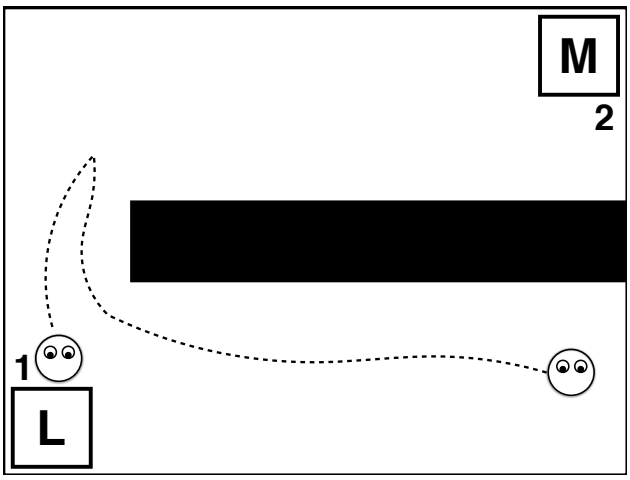


Figure 10: Food truck scenario in which an agent chooses which food truck to go to for lunch. *Note*: Numbers 1 and 2 indicate the two possible parking spots. M = Mexican food truck, L = Lebanese food truck. The agents' view of which truck is parked at parking spot 2 is blocked by a wall. The dotted line indicates the actual path that the agent took. Figure adapted from Baker et al. (2011).

al. (2011) have shown that people can infer the agent's preferences and beliefs merely based on the path that the agent walked. From the path in Figure 10, we can infer the agent's complete preference order for the three trucks. He likes the Korean truck best, and the Lebanese truck more than the Mexican truck. We can explain the agent's peeking around the corner by referring to his belief that the Korean truck might have been at parking spot 2. The principle of rational action implies that if the agent had known that the Korean truck wasn't parked at spot 2, he wouldn't have put in the effort to look around the corner. Instead, he would have directly gone for the Lebanese truck. Thus, in analogy to the causal judgments in the physical domain, we can explain other people's behavior in terms of counterfactual contrasts over our intuitive theory of psychology. Within this framework, we have already shown that people's attributions of responsibility are closely linked to their causal understanding of the situation (Gerstenberg, Halpern, & Tenenbaum, 2015; Gerstenberg & Lagnado, 2010, 2012; Lagnado et al., 2013; Zultan, Gerstenberg, & Lagnado, 2012) and their intuitive theory of how other people would (or should) have acted in a given situation (Allen, Jara-Ettinger, Gerstenberg, Kleiman-Weiner, & Tenenbaum, 2015; Gerstenberg, Ullman, Kleiman-Weiner, Lagnado, & Tenenbaum, 2014).

5 Conclusion and future directions

We started off this chapter with some of the big questions that were motivated by children's curiosity to figure out how the world works. Children rapidly develop an understanding of the world that is far beyond what can be captured by current approaches in artificial intelligence. Bridging the gap between human common-sense reasoning and machine intelligence requires acknowledging that people's knowledge of the world is structured in terms of intuitive theories (Forbus, 1984; Gopnik & Wellman, 1992; Saxe, 2005; Wellman & Gelman, 1992), and that many cognitive functions can be understood as inferences over these intuitive theories. We have argued that intuitive theories are best represented in terms of probabilistic, generative programs (Gerstenberg & Goodman, 2012; Goodman et al., 2015). We have provided empirical evidence for how understanding intuitive theories in terms of probabilistic, generative models allows to make sense of a wide array of cognitive phenomena (Chater & Oaksford, 2013; Danks, 2014). Because our intuitive theories are structured and generative, they support prediction, inference, action, counterfactual reasoning, and explanation for infinitely many possible situations.

Focusing on people's intuitive theory of physics and psychology, we have shown how people's causal judgments can be understood in terms of counterfactual contrasts defined over their intuitive understanding of the domain. Conceptualizing causal judgments in this way provides a bridge between process and dependency accounts of causation. Our proposed *counterfactual simulation model* accurately captures people's causal judgments about colliding billiard balls for a host of different situations, including interactions between two and three billiard balls with additional objects such as bricks. People's inferences about another agent's goals or intentions can be explained by assuming that we have an intuitive theory that others plan and make decisions in a rational manner.

The process of mental simulation plays a central role in this framework (Barsalou, 2009; Hegarty, 1992, 2004; Kahneman & Tversky, 1982; Wells & Gavanski, 1989; Yates et al., 1988). It provides the glue between people's abstract intuitive theories and the concrete inferences that are supported in a given situation through conditioning on what was observed (Battaglia et al., 2013) and imagining how things might have turned out differently (Gerstenberg et al., 2012; Gerstenberg, Goodman, et al., 2014, 2015). In the domain of intuitive physics, we have seen that people's predictions are

consistent with a noisy Newtonian framework that captures people's uncertainty by assuming that our mental simulations are guided by the laws of physics but that we are often uncertain about some aspects of the situation. Future research needs to study the process of mental simulation more closely and investigate what determines the quality and resolution of people's mental simulations (Crespi et al., 2012; Hamrick & Griffiths, 2014; Hamrick et al., 2015; Marcus & Davis, 2013; Schwartz & Black, 1996; Smith, Dechter, et al., 2013).

One of the key challenges for the line of work discussed in this chapter is to understand how people come to develop their intuitive understanding of how the world works (Friedman, Taylor, & Forbus, 2009). What are we endowed with from the start, and how do our representations of the world change over time (Carey, 2009)? How can we best model the process of theory acquisition (Gopnik, 2010)? We have seen above that the development of an intuitive theory of mind undergoes qualitatively different stages (Gopnik & Wellman, 1992) from an early theory that only considers goals and perceptual access (Gergely & Csibra, 2003) to a full-fledged theory of mind that integrates beliefs, desires, and intentions (Bratman, 1987; Malle, 1999). We have suggested that people's intuitive domain theories are best understood in terms of probabilistic, generative programs (Goodman et al., 2015). This raises the question of how such representations are learned (Tenenbaum et al., 2011). Computationally, this is known as the problem of program induction: learning a generative program based on data (e.g. Dechter, Malmaud, Adams, & Tenenbaum, 2013; Liang, Jordan, & Klein, 2010; Rule, Dechter, & Tenenbaum, 2015). Program induction is difficult since it is severely underconstrained: an infinite number of programs are consistent with any given data. Nevertheless, recent work has demonstrated that the problem is feasible. Different empirical phenomena such as number learning (Piantadosi, Tenenbaum, & Goodman, 2012), concept learning (Goodman, Tenenbaum, et al., 2008; Stuhlmüller, Tenenbaum, & Goodman, 2010), or acquiring a theory of causality (Goodman, Ullman, & Tenenbaum, 2011) have been cast as learning an intuitive theory through searching a hypothesis space of different possible programs that might have generated the data. This work has demonstrated how qualitative transitions in people's knowledge can be explained in terms of transitions between programs of different complexity (Piantadosi et al., 2012). While some first attempts have been made (Fragkiadaki, Agrawal, Levine, & Malik, submitted; Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2014), further work is required to explain how people arrive at their rich intuitive theories of how the world works.

6 Acknowledgments

This work was supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216 and by an ONR grant N00014-13-1-0333.

References

Abelson, R. P., & Lalljee, M. (1988). *Knowledge structures and causal explanation.* New York: New York University Press.

Ahn, W.-K., & Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. In F. Keil & R. Wilson (Eds.), *Explanation and cognition.* Cambridge, MA: Cambridge University Press.

Ahn, W.-K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*(3), 299–352.

Ajzen, I., & Fishbein, M. (1975). A Bayesian analysis of attribution processes. *Psychological Bulletin*, *82*(2), 261–277.

Ajzen, I., & Fishbein, M. (1983). Relevance and availability in the attribution process. In J. M. Jaspars, F. D. Fincham, & M. Hewstone (Eds.), *Advances in experimental social psychology* (pp. 63–89). New York: Academic Press.

Allen, K., Jara-Ettinger, J., Gerstenberg, T., Kleiman-Weiner, M., & Tenenbaum, J. B. (2015). Go fishing! responsibility judgments when cooperation breaks down. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 84–89). Austin, TX: Cognitive Science Society.

Apperly, I. A., Warren, F., Andrews, B. J., Grant, J., & Todd, S. (2011). Developmental continuity in theory of mind: Speed and accuracy of belief–desire reasoning in children and adults. *Child development*, *82*(5), 1691–1703.

Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, *13*(3), 89–94.

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, *20*(3), 191–208.

Baker, C. L., Goodman, N. D., & Tenenbaum, J. B. (2008). Theory-based social goal inference. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1447–1452).

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

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the*

33rd Annual Conference of the Cognitive Science Society. Austin, TX: : Cognitive Science Society.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

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.

Barsalou, L. W. (2009, mar). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1281–1289. Retrieved from `http://dx.doi.org/10.1098/rstb.2008.0319` doi: 10.1098/rstb.2008.0319

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Beebee, H., Hitchcock, C., & Menzies, P. (2009). *The oxford handbook of causation.* Oxford University Press, USA.

Birch, S. A., & Bloom, P. (2004). Understanding children's and adults' limitations in mental state reasoning. *Trends in Cognitive Sciences*, *8*(6), 255–260.

Birch, S. A., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*(5), 382–386.

Borkenau, P. (1992). Implicit personality theory and the five-factor model. *Journal of Personality*, *60*(2), 295–327.

Bowerman, W. R. (1978). Subjective competence: The structure, process and function of self-referent causal attributions. *Journal for the Theory of Social Behaviour*, *8*(1), 45–75.

Bramley, N., Gerstenberg, T., & Lagnado, D. A. (2014). The order of things: Inferring causal structure from temporal patterns. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 236–241). Austin, TX: Cognitive Science Society.

Bratman, M. (1987). *Intention, plans, and practical reason.* Center for the Study of Language and Information.

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind &*

Language, *28*(5), 606–637. Retrieved from http://dx.doi.org/10.1111/mila.12036 doi: 10.1111/mila.12036

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Chang, W. (2009). Connecting counterfactual and physical causation. In *Proceedings of the 31th annual conference of the cognitive science society* (pp. 1983–1987). Cognitive Science Society, Austin, TX.

Chater, N., & Oaksford, M. (2013). Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science*, *37*(6), 1171–1191.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40*, 83–120.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*(2), 365–382. Retrieved from http://dx.doi.org/10.1037/0033-295x.99.2.365 doi: 10.1037/0033-295x.99.2.365

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, *22*(1), 93–115.

Christensen-Szalanski, J. J., & Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational behavior and human decision processes*, *48*(1), 147–168.

Crespi, S., Robino, C., Silva, O., & de'Sperati, C. (2012). Spotting expertise in the eyes: Billiards knowledge as revealed by gaze shifts in a dynamic visual prediction task. *Journal of Vision*, *12*(11), 1–19.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153.

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. Mit Press.

Dechter, E., Malmaud, J., Adams, R. P., & Tenenbaum, J. B. (2013). Bootstrap learning via modular concept discovery. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence* (pp. 1302–1309).

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

De Vreese, L. (2006). Pluralism in the philosophy of causation: desideratum or not? *Philosophica*, *77*, 5–13.

DiSessa, A. A. (1982). Unlearning aristotelian physics: A study of knowledge-based learning. *Cogni-*

tive Science, *6*(1), 37–75. Retrieved from http://dx.doi.org/10.1207/s15516709cog0601_2 doi: 10.1207/s15516709cog0601_2

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, *10*(2-3), 105–225. Retrieved from http://dx.doi.org/10.1080/07370008.1985.9649008 doi: 10.1080/07370008.1985.9649008

Dowe, P. (2000). *Physical causation.* Cambridge University Press.

Downing, C. J., Sternberg, R. J., & Ross, B. H. (1985). Multicausal inference: Evaluation of evidence in causally complex situations. *Journal of Experimental Psychology: General*, *114*(2), 239–263.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*(1), 3–19.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327–339.

Fischhoff, B. (1976). Attribution theory and judgment under uncertainty. In J. H. Harvey, W. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 1). Hillsdale, NJ: Erlbaum.

Fischhoff, B., & Lichtenstein, S. (1978). Don't attribute this to reverend bayes. *Psychological Bulletin*, *85*(2), 239–243.

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology*, *50*(1), 21–45.

Fodor, J. A. (1975). *The language of thought.* Harvard University Press.

Forbus, K. D. (1984). Qualitative process theory. In *Qualitative reasoning about physical systems* (pp. 85–168). Elsevier BV. Retrieved from http://dx.doi.org/10.1016/b978-0-444-87670-6.50006-6 doi: 10.1016/b978-0-444-87670-6.50006-6

Forbus, K. D. (1993). Qualitative process theory: twelve years after. *Artificial Intelligence*, *59*(1), 115–123.

Forbus, K. D. (2010, sep). Qualitative modeling. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(4), 374–391. Retrieved from http://dx.doi.org/10.1002/wcs.115 doi: 10.1002/wcs.115

Fragkiadaki, K., Agrawal, P., Levine, S., & Malik, J. (submitted). Learning visual predictive models of physics for playing billiards. In *International conference on learning representations.*

Frazier, B. N., Gelman, S. A., & Wellman, H. M. (2009). Preschoolers' search for explanatory information within adult–child conversation. *Child development*, *80*(6), 1592–1611.

Friedman, S., Taylor, J., & Forbus, K. D. (2009). Learning naive physics models by analogical generalization. In B. Kokinov, K. Holyoak, & D. Gentner (Eds.), *Proceedings of the second international analogy conference* (pp. 145–154). Sofia, Bulgaria: NBU Press.

Gelman, S. A., & Legare, C. H. (2011, oct). Concepts and folk theories. *Annual Review of Anthropology*, *40*(1), 379–398. Retrieved from `http://dx.doi.org/10.1146/annurev-anthro-081309-145822` doi: 10.1146/annurev-anthro-081309-145822

Gentner, D. (2002). Psychology of mental models. In N. J. Smelser & P. B. Bates (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 9683–9687). Amsterdam: Elsevier Science.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Gerstenberg, T., & Goodman, N. D. (2012). Ping Pong in Church: Productive use of concepts in human probabilistic inference. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1590–1595). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 378–383). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2014). From counterfactual simulation to causal judgment. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 523–528). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787). Austin, TX:

Cognitive Science Society.

Gerstenberg, T., Halpern, J. Y., & Tenenbaum, J. B. (2015). Responsibility judgments in voting scenarios. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 788–793). Austin, TX: Cognitive Science Society.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*(1), 166–171.

Gerstenberg, T., & Lagnado, D. A. (2012). When contributions make a difference: Explaining order effects in responsibility attributions. *Psychonomic Bulletin & Review*, *19*(4), 729–736.

Gerstenberg, T., Ullman, T. D., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2014). Wins above replacement: Responsibility attributions as counterfactual replacements. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2263–2268). Austin, TX: Cognitive Science Society.

Glymour, C. N. (2001). *The mind's arrow: Bayes nets and graphical causal models.* MIT Press.

Godfrey-Smith, P. (2009). Causal pluralism. In H. Beebee, C. Hitchcock, & P. Menzies (Eds.), *Oxford handbook of causation* (pp. 326–337). Oxford University Press, USA.

Goldman, A. I. (2006). *Simulating minds.* Oxford University Press.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*(4), 565–610.

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., ... Tenenbaum, J. B. (2006). Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1382–1387).

Goodman, N. D., Baker, C. L., & Tenenbaum, J. B. (2009). Cause and intent: Social reasoning in causal learning. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society.*

Goodman, N. D., Mansinghka, V. K., Roy, D., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: A language for generative models. In *Uncertainty in Artificial Intelligence.*

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. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Lawrence (Eds.), *The conceptual mind: New directions in*

the study of concepts (pp. 623–653). MIT Press.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110.

Gopnik, A. (2010). How babies think. *Scientific American*, *303*(1), 76–81.

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, *337*(6102), 1623–1627.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. *Psychological Review*, *111*(1), 3–32.

Gopnik, A., & Wellman, H. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the Theory Theory. *Psychological Bulletin*, *138*(6), 1085–1108.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, *7*(1-2), 145–171.

Gordon, R. M. (1986). Folk psychology as simulation. *Mind & Language*, *1*(2), 158–171.

Gordon, R. M. (1992). The simulation theory: Objections and misconceptions. *Mind & Language*, *7*(1-2), 11–34.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716.

Gweon, H., & Schulz, L. E. (2011). 16-month-olds rationally infer causes of failed actions. *Science*, *332*(6037), 1524.

Hagmayer, Y., & Osman, M. (2012). From colliding billiard balls to colluding desperate housewives: causal bayes nets as rational models of everyday causal reasoning. *Synthese*, 1–12.

Hagmayer, Y., & Sloman, S. A. (2009). Decision makers conceive of their choices as interventions. *Journal of Experimental Psychology: General*, *138*(1), 22–38.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals.* MIT Press.

Halpern, J. Y. (2008). Defaults and normality in causal structures. In *Proceedings of the 11th Conference on Principles of Knowledge Representation and Reasoning* (pp. 198–208).

Halpern, J. Y., & Hitchcock, C. (forthcoming). Graded causation and defaults.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, *56*(4), 843–887.

Hamrick, J. B., & Griffiths, T. L. (2014). What to simulate? inferring the direction of mental rotation. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? the amount of mental simulation tracks uncertainty in the outcome. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 866–871). Austin, TX: Cognitive Science Society.

Hartshorne, J. K. (2013, may). What is implicit causality? *Language, Cognition and Neuroscience*, *29*(7), 804–824. Retrieved from `http://dx.doi.org/10.1080/01690965.2013.796396` doi: 10.1080/01690965.2013.796396

Harvey, J. H., Ickes, W. J., & Kidd, R. F. (1978). *New directions in attribution research* (Vol. 2). Lawrence Erlbaum Associates.

Hayes, P. J. (1985). The second naive physics manifesto. In J. Hobbs & R. Moore (Eds.), .

Hegarty, M. (1992). Mental animation: inferring motion from static displays of mechanical systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 1084–1102.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

Heider, F. (1958). *The psychology of interpersonal relations.* John Wiley & Sons Inc.

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

Henderson, L., Goodman, N. D., Tenenbaum, J. B., & Woodward, J. F. (2010). The structure and dynamics of scientific theories: A hierarchical bayesian perspective. *Philosophy of Science*, *77*(2), 172–200.

Hewstone, M., & Jaspars, J. (1987). Covariation and causal attribution: A logical model of the intuitive analysis of variance. *Journal of Personality and Social Psychology*, *53*(4), 663.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*,

107(1), 65–81.

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, *93*(1), 75–88.

Hitchcock, C. (2009). Structural equations and causation: six counterexamples. *Philosophical Studies*, *144*(3), 391–401.

Hume, D. (1748/1975). *An enquiry concerning human understanding.* Oxford University Press.

Jara-Ettinger, J., Baker, C. L., & Tenenbaum, J. B. (2012). Learning what is where from social observations. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015, jul). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23. Retrieved from `http://dx.doi.org/10.1016/j.cognition.2015.03.006` doi: 10.1016/j.cognition.2015.03.006

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not so innocent: Toddlers' iinferences about costs and culpability. *Psychological Science*, *26*(5), 633–640. Retrieved from `http://dx.doi.org/10.1177/0956797615572806` doi: 10.1177/0956797615572806

Jaspars, J., Hewstone, M., & Fincham, F. D. (1983). Attribution theory and research: The state of the art. In J. M. Jaspars, F. D. Fincham, & M. Hewstone (Eds.), *Attribution theory and research: Conceptual, developmental and social dimensions* (pp. 343–369). New York: Academic Press.

Jaspars, J. M., Fincham, F. D., & Hewstone, M. (1983). *Attribution theory and research: Conceptual, developmental and social dimensions.* New York: Academic Press.

Jonze (Director), S. (2013). *Her.* Annapurna Pictures. United States.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge University Press.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). New York: Cambridge University Press.

Kaiser, M. K., Proffitt, D. R., & McCloskey, M. (1985). The development of beliefs about falling objects. *Attention, Perception, & Psychophysics*, *38*(6), 533–539.

Kaiser, M. K., Proffitt, D. R., Whelan, S. M., & Hecht, H. (1992). Influence of animation on dynamical judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 669–690.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences*, *7*(8), 368–373.

Keil, F. C. (2012, oct). Running on Empty? How Folk Science Gets By With Less. *Current Directions in Psychological Science*, *21*(5), 329–334. Retrieved from http://dx.doi.org/10.1177/0963721412453721 doi: 10.1177/0963721412453721

Kelley, H. H. (1972). *Causal schemata and the attribution process.* New York: General Learning Press.

Kleer, J. D., & Brown, J. S. (1984). A qualitative physics based on confluences. In *Qualitative reasoning about physical systems* (pp. 7–83). Elsevier. Retrieved from http://dx.doi.org/10.1016/b978-0-444-87670-6.50005-4 doi: 10.1016/b978-0-444-87670-6.50005-4

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1123–1128). Austin, TX: Cognitive Science Society.

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.

Kozhevnikov, M., & Hegarty, M. (2001). Impetus beliefs as default heuristics: Dissociation between explicit and implicit knowledge about motion. *Psychonomic Bulletin & Review*, *8*(3), 439–453.

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, *136*(3), 430–450.

Kuhn, T. S. (1996). *The structure of scientific revolutions.* University of Chicago Press. Retrieved from http://dx.doi.org/10.7208/chicago/9780226458106.001.0001 doi: 10.7208/chicago/9780226458106.001.0001

Kuhnmünch, G., & Beller, S. (2005, Nov). Distinguishing between causes and enabling conditions-through mental models or linguistic cues? *Cognitive Science*, *29*(6), 1077-1090. Retrieved from http://dx.doi.org/10.1207/s15516709cog0000_39 doi: 10.1207/s15516709cog0000

_39

Lagnado, D. A. (2011). Causal thinking. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 129–149). Oxford: Oxford University Press.

Lagnado, D. A., Fenton, N., & Neil, M. (2012). Legal idioms: A framework for evidential reasoning. *Argument and Computation*.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *47*, 1036–1073.

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856–876.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 451–460.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). Oxford University Press.

Levesque, H. J., Davis, E., & Morgenstern, L. (2011). The winograd schema challenge. In *Aaai spring symposium: Logical formalizations of commonsense reasoning.*

Levillain, F., & Bonatti, L. L. (2011). A dissociation between judged causality and imagined locations in simple dynamic scenes. *Psychological science*, *22*(5), 674–681.

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*(17), 556–567.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*(4), 455–476.

Lewis, D. (1986). Postscript C to 'Causation': (Insensitive causation). In *Philosophical papers* (Vol. 2). Oxford: Oxford University Press.

Lewis, D. (2000). Causation as influence. *The Journal of Philosophy*, *97*(4), 182–197.

Liang, P., Jordan, M. I., & Klein, D. (2010). Learning programs: A hierarchical bayesian approach. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (pp. 639–646).

Lipe, M. G. (1991). Counterfactual reasoning as a framework for attribution theories. *Psychological Bulletin*, *109*(3), 456–471.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332.

Lombrozo, T. (2012). *Explanation and abductive inference.* Oxford University Press. Retrieved from http://dx.doi.org/10.1093/oxfordhb/9780199734689.013.0014 doi: 10.1093/oxfordhb/9780199734689.013.0014

Lucas, C., Griffiths, T. L., Xu, F., & Fawcett, C. (2009). A rational model of preference learning and choice prediction by children. *Advances in Neural Information Processing Systems*, *21*, 985–992.

Mackie, J. L. (1974). *The cement of the universe.* Oxford: Clarendon Press.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, *3*(1), 23–48.

Malle, B. F. (2008). Fritz heider's legacy: Celebrated insights, many of them misunderstood. *Social Psychology*, *39*(3), 163–173.

Malle, B. F. (2011). Time to give up the dogmas of attribution: An alternative theory of behavior explanation. *Advances in Experimental Social Psychology*, *44*, 297–352.

Malle, B. F., Knobe, J., O'Laughlin, M. J., Pearce, G. E., & Nelson, S. E. (2000). Conceptual structure and social functions of behavior explanations: Beyond person–situation attributions. *Journal of Personality and Social Psychology*, *79*(3), 309–326.

Mandel, D. R. (2003). Judgment dissociation theory: An analysis of differences in causal, counterfactual and covariational reasoning. *Journal of Experimental Psychology: General*, *132*(3), 419–434.

Marcus, G. F., & Davis, E. (2013, oct). How robust are probabilistic models of higher-level cognition? *Psychological Science*, *24*(12), 2351–2360. Retrieved from http://dx.doi.org/10.1177/0956797613495418 doi: 10.1177/0956797613495418

McCloskey, M. (1983). Naive theories of motion. In *Mental models* (pp. 299–324).

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naíve beliefs about the motion of objects. *Science*, *210*(4474), 1138–1141.

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 636–649.

Meder, B., Gerstenberg, T., Hagmayer, Y., & Waldmann, M. R. (2010). Observing and intervening: Rational and heuristic models of causal decision making. *Open Psychology Journal*, *3*, 119–

135.

Michotte, A. (1946/1963). *The perception of causality*. Basic Books.

Mills, C. M., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology*, *87*(1), 1–32.

Mischel, W. (2004). Toward an integrative science of the person. *Annual Review of Psychology*, *55*, 1–22.

Mitchell, J. P. (2006, mar). Mentalizing and marr: An information processing approach to the study of social cognition. *Brain Research*, *1079*(1), 66–75. Retrieved from `http://dx.doi.org/10.1016/j.brainres.2005.12.113` doi: 10.1016/j.brainres.2005.12.113

Mochon, D., & Sloman, S. A. (2004). Causal models frame interpretation of mathematical equations. *Psychonomic Bulletin & Review*, *11*(6), 1099–1104. Retrieved from `http://dx.doi.org/10.3758/bf03196743` doi: 10.3758/bf03196743

Muentener, P., Friel, D., & Schulz, L. (2012, aug). Giving the giggles: Prediction, intervention, and young children's representation of psychological events. *PloS One*, *7*(8), e42495. Retrieved from `http://dx.doi.org/10.1371/journal.pone.0042495` doi: 10.1371/journal.pone.0042495

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, *65*(3), 151–166.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Ojalehto, B., Waxman, S. R., & Medin, D. L. (2013). Teleological reasoning about nature intentional design or relational perspectives. *Trends in Cognitive Sciences, in press*.

Paul, L. A., & Hall, N. (2013). *Causation: A user's guide*. Oxford University Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

Pearl, J. (1999). Probabilities of causation: three counterfactual interpretations and their identification. *Synthese*, *121*(1-2), 93–149.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge University Press.

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, *5*(2), 125–137.

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), 199–217.

Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition*, *36*(1), 1–16.

Rottman, B. M., & Hastie, R. (2013). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*. Retrieved from `http://dx.doi.org/10.1037/a0031903` doi: 10.1037/a0031903

Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: observations and interventions. *Cognitive psychology*, *64*(1), 93–125.

Rule, J., Dechter, E., & Tenenbaum, J. B. (2015). Representing and learning a large system of number concepts with latent predicate networks. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2051–2056). Austin, TX: Cognitive Science Society.

Rumelhart, D. E., & McClelland, J. L. (1988). *Parallel distributed processing.* MIT Press.

Sagi, E., & Rips, L. J. (2014, aug). Identity, causality, and pronoun ambiguity. *Topics in Cognitive Science*, *6*(4), 663–680. Retrieved from `http://dx.doi.org/10.1111/tops.12105` doi: 10.1111/tops.12105

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world.* Princeton University Press, Princeton NJ.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, *120*(2), 411–437.

Saxe, R. (2005). Against simulation: the argument from error. *Trends in cognitive sciences*, *9*(4), 174–179.

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, *55*, 87–124.

Saxe, R., Tenenbaum, J., & Carey, S. (2005, dec). Secret agents: Inferences about hidden causes by 10- and 12-month-old infants. *Psychological Science*, *16*(12), 995–1001. Retrieved from `http://dx.doi.org/10.1111/j.1467-9280.2005.01649.x` doi: 10.1111/j.1467-9280.2005.01649.x

Schächtele, S., Gerstenberg, T., & Lagnado, D. A. (2011). Beyond outcomes: The influence of intentions and deception. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1860–1865). Austin, TX: Cognitive Science Society.

Schaffer, J. (2005). Contrastive causation. *The Philosophical Review*, *114*(3), 327–358.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding: An inquiry into human knowledge structures.* Hillsdale, NJ: Erlbaum.

Schlottmann, A. (1999). Seeing it happen and knowing how it works: How children understand the relation between perceptual causality and underlying mechanism. *Developmental psychology*, *35*, 303–317.

Schulz, L. (2012). The origins of inquiry: inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, *16*(7), 382–389.

Schulz, L. E., Goodman, N. D., Tenenbaum, J. B., & Jenkins, A. C. (2008). Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition*, *109*(2), 211–223.

Schwartz, D. L., & Black, J. B. (1996, apr). Analog imagery in mental model reasoning: Depictive models. *Cognitive Psychology*, *30*(2), 154–219. Retrieved from `http://dx.doi.org/10.1006/cogp.1996.0006` doi: 10.1006/cogp.1996.0006

Scott, R. M., & Baillargeon, R. (2013). Do infants really expect agents to act efficiently? a critical test of the rationality principle. *Psychological Science, in press*.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others the consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*(4), 341–351.

Shanks, D. R., & Dickinson, A. (1987). Associative accounts of causality judgment. In *The psychology of learning and motivation* (Vol. 21, pp. 229–261). San Diego, CA: Academic Press.

Shanon, B. (1976). Aristotelianism, newtonianism and the physics of the layman. *Perception*, *5*(2), 241–243.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, 1–51.

Simons, D. J. (2000). Attentional capture and inattentional blindness. *Trends in cognitive sciences*, *4*(4), 147–155.

Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives.* Oxford University Press, USA.

Sloman, S. A., Barbey, A. K., & Hotaling, J. M. (2009). A causal model theory of the meaning of cause, enable, and prevent. *Cognitive Science*, *33*(1), 21–50.

Sloman, S. A., Fernbach, P. M., & Ewing, S. (2012). A causal model of intentionality judgment. *Mind and Language*, *27*(2), 154–180.

Sloman, S. A., & Hagmayer, Y. (2006). The causal psycho-logic of choice. *Trends in Cognitive Sciences*, *10*(9), 407–412.

Sloman, S. A., & Lagnado, D. A. (2005). Do we 'do'? *Cognitive Science*, *29*(1), 5–39.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology*, *28*(1), 1–39.

Smith, K. A., Battaglia, P., & Vul, E. (2013). Consistent physics underlying ballistic motion prediction. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Smith, K. A., Dechter, E., Tenenbaum, J., & Vul, E. (2013). Physical predictions over time. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the cognitive science society* (pp. 1342–1347).

Smith, K. A., & Vul, E. (2012). Sources of uncertainty in intuitive physics. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, *5*(1), 185–199.

Smith, K. A., & Vul, E. (2014). Looking forwards and backwards: Similarities and differences in prediction and looking forwards and backwards: Similarities and differences in prediction and retrodiction. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1467–1472). Austin, TX: Cognitive Science Society.

Sodian, B., Schoeppner, B., & Metz, U. (2004). Do infants apply the principle of rational action

to human agents? *Infant Behavior and Development*, *27*(1), 31–41.

Spelke, E. S. (1990, jan). Principles of object perception. *Cognitive Science*, *14*(1), 29–56. Retrieved from http://dx.doi.org/10.1207/s15516709cog1401_3 doi: 10.1207/s15516709cog1401_3

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*(4), 605–632. Retrieved from http://dx.doi.org/10.1037/0033-295X.99.4.605 doi: 10.1037/0033-295x.99.4.605

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489.

Stich, S., & Nichols, S. (1992). Folk psychology: Simulation or tacit theory? *Mind & Language*, *7*(1-2), 35–71.

Strevens, M. (2013, jul). Causality reunified. *Erkenntnis*, *78*(S2), 299–320. Retrieved from http://dx.doi.org/10.1007/s10670-013-9514-8 doi: 10.1007/s10670-013-9514-8

Stuhlmüller, A., Tenenbaum, J. B., & Goodman, N. D. (2010). Learning structured generative concepts. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Suppes, P. (1970). *A probabilistic theory of causation.* North-Holland.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*(1), 49–100.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, *332*(6033), 1054–1059.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In *Causal learning: Psychology, philosophy, and computation.* Oxford University Press.

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.

Todorov, E. (2004, sep). Optimality principles in sensorimotor control. *Nature Neuroscience*, *7*(9), 907–915. Retrieved from http://dx.doi.org/10.1038/nn1309 doi: 10.1038/nn1309

Tremoulet, P. D., & Feldman, J. (2006). The influence of spatial context and the role of intention-

ality in the interpretation of animacy from motion. *Attention, Perception, & Psychophysics*, *68*(6), 1047–1058.

Tremoulet, P. D., Feldman, J., et al. (2000). Perception of animacy from the motion of a single object. *Perception*, *29*(8), 943–952.

Uleman, J. S., Adil Saribay, S., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology*, *59*, 329–360.

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2014). Learning physics from dynamical scenes. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society.*

Ullman, T. D., Tenenbaum, J. B., Baker, C. L., Macindoe, O., Evans, O. R., & Goodman, N. D. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in Neural Information Processing Systems* (Vol. 22).

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 216–227.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*(2), 222–236.

Walsh, C. R., & Sloman, S. A. (2011). The meaning of cause and prevent: The role of causal mechanism. *Mind & Language*, *26*(1), 21–52.

Wellman, H. M. (2011). The child's theory of mind.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child development*, *72*(3), 655–684.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*(1), 337–375.

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology*, *56*(2), 161–169.

White, P. A. (1990). Ideas about causation in philosophy and psychology. *Psychological Bulletin*, *108*(1), 3–18. Retrieved from `http://dx.doi.org/10.1037/0033-2909.108.1.3` doi: 10 .1037/0033-2909.108.1.3

White, P. A. (2006). The causal asymmetry. *Psychological Review*, *113*(1), 132–147. Retrieved from http://dx.doi.org/10.1037/0033-295x.113.1.132 doi: 10.1037/0033-295x.113.1.132

White, P. A. (2009). Perception of forces exerted by objects in collision events. *Psychological Review*, *116*(3), 580–601.

White, P. A. (2012a, aug). The impetus theory in judgments about object motion: A new perspective. *Psychonomic Bulletin & Review*, *19*(6), 1007–1028. Retrieved from http://dx.doi.org/10.3758/s13423-012-0302-2 doi: 10.3758/s13423-012-0302-2

White, P. A. (2012b). Visual impressions of causality: Effects of manipulating the direction of the target object's motion in a collision event. *Visual Cognition*, *20*(2), 121–142.

Williamson, J. (2006). Causal pluralism versus epistemic causality. *Philosophica*, *77*, 69–96.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Wolff, P. (2003). Direct causation in the linguistic coding and individuation of causal events. *Cognition*, *88*(1), 1–48.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.

Wolff, P., Barbey, A. K., & Hausknecht, M. (2010). For want of a nail: How absences cause events. *Journal of Experimental Psychology: General*, *139*(2), 191–221.

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford, England: Oxford University Press.

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, *115*(1), 1–50.

Woodward, J. (2011). Mechanisms revisited. *Synthese*, *183*(3), 409–427.

Yates, J., Bessman, M., Dunne, M., Jertson, D., Sly, K., & Wendelboe, B. (1988). Are conceptions of motion based on a naive theory or on prototypes? *Cognition*, *29*(3), 251–275.

Ybarra, O. (2002). Naive causal understanding of valenced behaviors and its implications for social information processing. *Psychological Bulletin*, *128*(3), 421–441.

Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. *PLoS Computational Biology*, *4*(12), e1000254.

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science*, *28*(6), 979–1008.

Zago, M., & Lacquaniti, F. (2005, jan). Cognitive, perceptual and action-oriented representations of falling objects. *Neuropsychologia*, *43*(2), 178–188. Retrieved from `http://dx.doi.org/10.1016/j.neuropsychologia.2004.11.005` doi: 10.1016/j.neuropsychologia.2004.11.005

Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Counterfactuals and causality in group attributions. *Cognition*, *125*(3), 429–440.