Noisy Newtons:
Unifying process and dependency accounts of causal attribution

Tobias Gerstenberg[1] (t.gerstenberg@ucl.ac.uk), Noah Goodman[2] (ngoodman@stanford.edu),
David A. Lagnado[1] (d.lagnado@ucl.ac.uk) & Joshua B. Tenenbaum[3] (jbt@mit.edu)
[1]Cognitive, Perceptual and Brain Sciences, University College London, London WC1H 0AP
[2]Department of Psychology, Stanford University, Stanford, CA 94305
[3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Abstract

There is a long tradition in both philosophy and psychology to separate process accounts from dependency accounts of causation. In this paper, we motivate a unifying account that explains people's causal attributions in terms of counterfactuals defined over probabilistic generative models. In our experiments, participants see two billiard balls colliding and indicate to what extent ball A caused/prevented ball B to go through a gate. Our model predicts that people arrive at their causal judgments by comparing what actually happened with what they think would have happened, had the collision between A and B not taken place. Participants' judgments about what would have happened are highly correlated with a noisy model of Newtonian physics. Using those counterfactual judgments, we can predict participants' cause and prevention judgments very accurately ($r = .99$). Our framework also allows us to capture intrinsically counterfactual judgments such as *almost* caused/prevented.

Keywords: causality; counterfactuals; attribution; physics.

Introduction

There has been a longstanding divide in philosophy between two fundamentally different ways of conceptualizing causality. According to *dependency accounts* of causation, what it means for A to be a cause of B is that B is in some way dependent on A. Dependence has been conceptualized in terms of regularity of succession (A is regularly succeeded by B; Hume, 2000 [1748]), probabilities (the presence of A increases the probability of B; Suppes, 1970) or counterfactuals (if A had not been present B would not have occurred; Lewis, 1970). For *process accounts*, in contrast, what it means for A to be a cause of B is that a physical quantity is transmitted along a pathway from A to B (Dowe, 2000).

The psychological literature on causal learning and attribution neatly maps onto the two major accounts in philosophy. On the one hand, people have been shown to use contingency information when drawing inferences about whether and how strongly two events are causally linked (Cheng, 1997). On the other hand, people display a preference to choose causes that influence an effect via a continuous causal mechanism over causes that are connected with the effect through mere dependence (Walsh & Sloman, 2011).

A point that has been raised in favor of process accounts is that they are capable of capturing the semantics of different causal terms. Whereas dependency accounts have mostly focussed on causation and prevention, Wolff (2007) has provided a process account that not only predicts when people use the terms *cause* and *prevent* but also *enable* and *despite*. Following a linguistic analysis of causation by Talmy (1988) in terms of force dynamics, Wolff (2007) argues that the aforementioned causal terms can be reduced to configurations of force vectors. For example, what it means for a patient (P) to have been caused by an affector (A) to reach an endstate (E) is that P did not have a tendency towards E, A impacted on P in a way that their force vectors were not pointing in the same direction and P reached E. If, in contrast, the force vectors of both P and A point towards E and P reaches E, the model predicts that people will say "A enabled (rather than caused) P". Importantly, according to Wolff's account, the core dimensions which underlie the different causal terms, such as P's tendency towards E, are defined in strictly non-counterfactual terms. Hence, "tendency" is defined as the direction of P's force rather than whether P would reach E in the absence of any other forces.

While the force dynamics model has strong intuitive appeal for interactions between physical entities, it is questionable how it can be extended to capture causal attributions in situations involving more abstract entities. For example, one might legitimately assert that the fall of Lehman Brothers caused the financial crisis or that Tom's belief that he forgot his keys caused him to turn around and go back home. While it is unclear how these causal relationships could be expressed in terms of force vectors, they do not pose a problem for the more flexible dependency accounts. For example, according to a counterfactual account, Tom's belief qualifies as cause of his behaviour if it is true that his behavior would have been different had the content of his belief been different. Hence, there appears to be a trade-off between the semantic richness of process accounts on the one hand and the generality and flexibility of dependency accounts on the other hand.

Rather than fostering the divide between process accounts and dependency accounts, we propose a theory of causal attribution that combines the best of both worlds. In the spirit of Pearl (2000), we model causal attributions in terms of counterfactuals defined over probabilistic generative models. However, we agree with Wolff (2007) that people's causal knowledge is often richer than what can be expressed with a causal Bayes net. We aim to unify process and dependency accounts by showing that people have intuitive theories in the form of detailed generative models, and that causal judgements are made by considering counterfactuals over these intuitive theories. Here we demonstrate the superiority of our approach over existing models of causal attribution in a physical domain. We show that people use their intuitive understanding of physics to simulate possible future outcomes and that their causal attributions are a function of what actually happened and their belief about what would have happened had the cause not been present.

Overview of Experiments and Model Predictions

Before discussing the predictions of our model and the supporting evidence from four experiments, we describe the domain to which we applied our model. In all experiments, participants saw the same 18 video clips which were generated by implementing the physics engine Box2D into Adobe Flash CS5. Figure 1 depicts a selection of the clips.[1] In each clip, there was a single collision event between a grey ball (A) and a red ball (B) which enter the scene from the right. Collisions were elastic and there was no friction. The black bars are solid walls and the red bar on the left is a gate that balls can go through. In some clips B went through the gate (e.g. clip 18) while in others it did not (e.g. clip 5). In the 18 video clips, we crossed whether ball B went through the gate given that it collided with ball A (rows in Figure 1: actual miss/close/hit) with whether B would go through the gate if A was not present in the scene (columns in Figure 1: counterfactual miss/close/hit). Participants viewed two clips for each combination of actual and counterfactual outcome.

In Experiments 1 and 2, the video clips stopped shortly after the collision event. Participants judged whether ball B will go through the gate (Experiment 1) or whether ball B would go through the gate if ball A was not present in the scene (Experiment 2). In Experiment 3, participants saw each clip played until the end and then judged to what extent ball A caused ball B to go through the gate or prevented B from going through the gate. Finally, in Experiment 4 participants chose from a set of sentences which best describes the clip they have just seen. All experiments were run online and participants were recruited via Amazon Mechanical Turk.

In order to model people's predictions of actual and counterfactual future outcomes, we developed the Physics Simulation Model (PSM) which assumes that people make use of their intuitive understanding of physics to simulate what will or what might have happened. Hamrick, Battaglia, and Tenenbaum (2011) have shown that people's stability judgments about towers of blocks is closely in line with a noisy model of Newtonian physics. While in their model, the introduced noise captures people's uncertainty about the exact location of each block, the noise in our model captures the fact that people cannot perfectly predict the trajectory of a moving ball (cf. Figure 1, clip 1). We introduce noise via drawing different degrees of angular perturbation from a Gaussian distribution with $M = 0$ and $SD = \{1, 2, ..., 10\}$ which is then applied to B's actual velocity vector (given that it collided with A, clip 1 bottom left) or B's counterfactual velocity vector (given that A was not present in the scene, clip 1 top left) at the time of collision.

We evaluate the probability that B would go through the gate when A was present, $P(B|A)$, or absent $P(B|\neg A)$ by forward sampling from our noisy versions of Newtonian physics.

[1] All clips can be viewed here: http://www.ucl.ac.uk/lagnado-lab/experiments/demos/physicsdemo.html

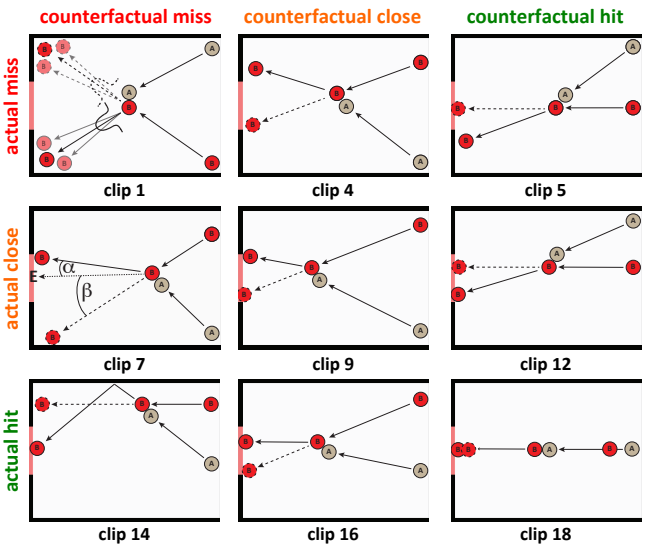

Figure 1: Selection of clips used in the experiment. Solid arrows = actual paths, dashed arrows = counterfactual paths. Clip 1 depicts an illustration of the Physics Simulation Model and clip 7 of the Actual Force Model. *Note*: actual miss = B clearly misses; actual close = B just misses/hits; actual hit = B clearly hits; counterfactual miss = B would have clearly missed; counterfactual close = B would have just missed/hit; counterfactual hit = B would have clearly hit.

For each clip and degree of noise (SD), we ran 1000 noisy repetitions of the original clip and counted the worlds in which B goes through the gate given that A was present $P(B|A)$ or absent $P(B|\neg A)$.

Experiments 1 & 2: Intuitive Physics

The aim of Experiments 1 and 2 was to evaluate how well people make use of their intuitive physical knowledge to predict actual (Experiment 1) or counterfactual (Experiment 2) future states. Participants saw 18 video clips (see Figure 1 for some examples) up to the point shortly after the two balls collided ($0.1s$). After having seen the clip twice, participants answered the question: "Will the red ball go through the hole?" (Experiment 1, $N = 21$) or "Would the red ball have gone through the goal if the gray ball had not been present?" (Experiment 2, $N = 20$). Participants indicated their answers on a slider that ranged from 0 ("definitely no") to 100 ("definitely yes"). The midpoint was labeled "uncertain". After having made their judgment, participants viewed the clip until the end either with both balls being present (Experiment 1) or with ball A being removed from the scene (Experiment 2).

Results and Discussion

Participants were accurate in judging whether ball B will go through the gate (Experiment 1) or would have gone through the gate (Experiment 2) with a mean absolute difference from the deterministic physics model (which assigns a value of 100 if B goes in and 0 if B does not go in) of 28.6 ($SD = 29.9$) in Experiment 1 and 25.1 ($SD = 30.5$) in Experiment 2. Figure 2 shows the correlation of the PSM with participants' judgments in Experiment 1 (solid black line) and Experiment 2 (dashed black line) for different degrees of noise. While people's judgments already correlate quite well with a deterministic Newto-

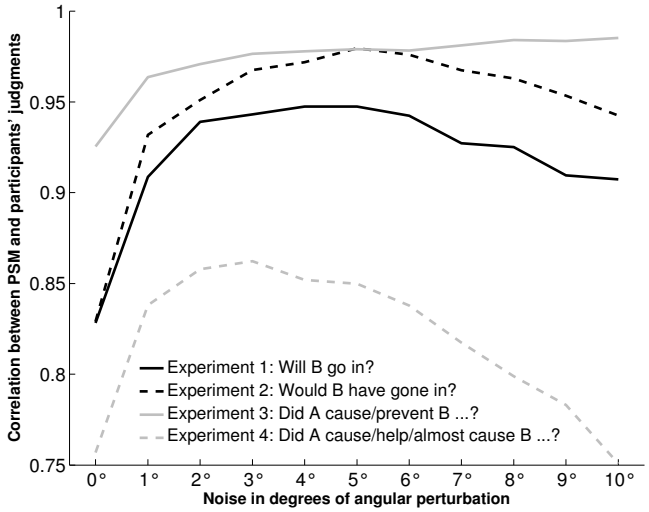

Figure 2: Correlation of the Physics Simulation Model with people's judgments in all four experiments for different degrees of noise.

nian physics model (degree of noise = 0°), introducing small degrees of noise results in much higher correlations with a maximum correlation of $r = .95$ in Experiment 1 and $r = .98$ in Experiment 2 for $SD = 5°$.

The results of Experiments 1 and 2 show that people are capable of mentally simulating what will happen (Experiment 1) or what would have happened (Experiment 2). Given that each clip stopped very shortly after the collision event, participants' accuracy in judging whether ball B will go in or not is quite impressive.

Experiment 3: Causation and Prevention

In Experiment 3, we wanted to investigate how people use their intuitive understanding of physics to make judgments about the extent to which one event caused or prevented another event from happening. Unlike in Experiments 1 and 2, participants (N = 22) saw each clip played until the end. After having seen each clip twice, participants answered the question "What role did ball A play?" by moving a slider whose endpoints were labeled with "it prevented B from going through the hole" and "it caused B to go through the hole". The midpoint was labeled "neither". The slider ranged from -100 (prevented) to 100 (caused). Participants were instructed that they could use intermediate values on the slider to indicate that ball A somewhat caused or prevented B.

Model Predictions

Physics Simulation Model According to the PSM, people arrive at their cause and prevention judgments by comparing what actually happened with what they think would have happened if the cause event had not taken place. More specifically, our model predicts that people compare $P(B|A)$, the probability that ball B will go through the gate given that it collided with ball A, with $P(B|\neg A)$, the probability that B would have gone through the gate if A had not been present in the scene. Since participants in Experiment 3 watch the clips until the end, the value of $P(B|A)$ is certain: it is either 1 when B goes through the gate or 0 when B misses the gate. In order to determine $P(B|\neg A)$, the PSM assumes that people use their confidence in the result of their mental simulation of what would have happened had A not been present.

In general, if $P(B|A) - P(B|\neg A)$ is negative, participants should say that A prevented B from going through the gate. Intuitively, if it was obvious that B would have gone in had A not been present (i.e. $P(B|\neg A)$ is high) but B misses the gate as a result of colliding with A (i.e. $P(B|A) = 0$), A should be judged to have prevented B from going through the gate. Similarly, if the difference is positive, participants should say that A caused B to go through the gate. If the chance that B would have gone through the goal without A was low but, as a result of colliding with A, B goes through the gate, A should be judged to have caused B to go through the gate. Clip 1 in Figure 1 shows an example for which our model predicts that participants will say that A neither caused nor prevented B. $P(B|A)$ is 0 since B does not go through the gate. However, $P(B|\neg A)$ is also close to 0 since it is clear that B would have missed the gate even if A had not been present in the scene.

Actual Force Model The Actual Force Model (AFM) is our best attempt to apply Wolff's (2007) force dynamics model to our task.[2] According to the AFM, participants' cause and prevention judgments are a direct result of the physical forces which are present at the time of collision.

Clip 7 in Figure 1 illustrates how the AFM works. First, a goal vector (dotted arrow) is drawn from ball B's location at the time of collision to an end state (E), which we defined to be in the center of the gate. Second, the angle α between the velocity vector that ball B has shortly *after* the collision with A (solid arrow) and the goal vector as well as the angle β between the velocity vector that ball B has shortly *before* colliding with A (dashed arrow) are determined. Third, the model predicts people's cause and prevention judgments via comparison of α and β. In general, if ball B goes in and $\beta - \alpha$ is greater than 0, the model predicts people will say that A caused B. Conversely, if ball B does not go in and $\beta - \alpha$ is smaller than 0, the model predicts people will say A prevented B. For situations in which $\beta - \alpha$ is greater than 0 but B does not go in or $\beta - \alpha$ is smaller than 0 but B does go in, we fix the model prediction to 0. This constraint prevents the model from predicting, for example, that people will say "A caused B" when B missed the gate.

Results and Discussion

Figure 3 shows participants' mean cause and prevention judgments for the 18 different clips together with the predictions of the PSM and the AFM. For the particular implementation of the PSM depicted in Figure 3, we directly used participants' judgments from Experiment 2 in which they indicated whether ball B would have gone through the gate if ball A had not been present as the values for $P(B|\neg A)$. For example, in clip 5 the ball misses the gate (hence $P(B|A) = 0$) and

[2]While the force dynamics model only makes predictions about which out of several sentences participants will choose to describe a situation, the AFM makes quantitative predictions about the extent to which an event is seen as causal/preventive.

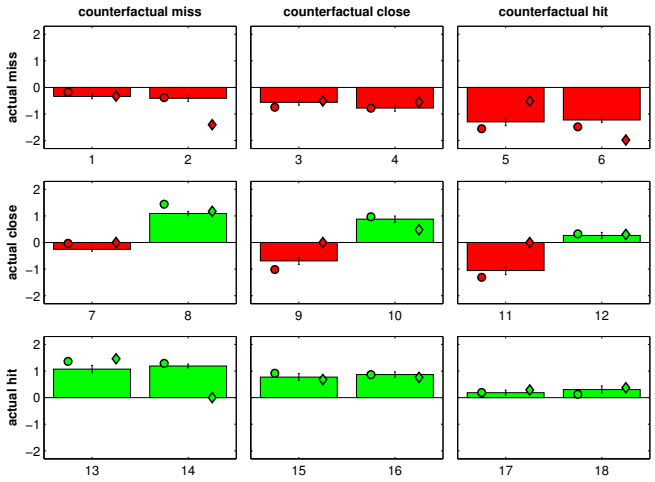

Figure 3: Z-scored mean cause (green) and prevention ratings (red) for the different clips denoted on the x-axes. ○ = predictions of the Physics Simulation Model ($r = .99$), ◇ = predictions of the Actual Force Model ($r = .77$). Error bars are ± 1 *SEM*.

participants' average confidence rating from Experiment 2 of whether B would have gone through in the absence of A is 97% (hence $P(B|\neg A) = .97$). Thus the PSM predicts that participants will indicate that A strongly prevented B in this clip, because $P(B|A) - P(B|\neg A)$ is close to the minimum of -1.

Overall, the PSM predicts participants' cause and prevention ratings very well with $r = .99$ and *RMSE* = 0.02. A high median correlation across participants of $r = .88$ with a minimum of $r = .61$ and a maximum of $r = .95$ demonstrates that the good performance of the PSM is not due to a mere aggregation effect. The PSM achieves its high predictive accuracy without the need for any free parameters. We directly used participants' judgments from Experiment 2 to determine the value of $P(B|\neg A)$ for each clip. Figure 2 shows that participants' judgments also correlate highly with the PSM when we generate $P(B|\neg A)$ through the noisy simulations of Newtonian physics as described above.

The AFM, in contrast, does not predict participants' judgments equally well with a correlation of $r = .77$ and *RMSE* = 0.44. While the AFM predicts people's judgments for many of the clips, there are a number of clips for which its predictions are inaccurate (most notably: clips 2, 5, 9, 11 and 14).

Interestingly, people's cause and prevention judgments were not affected by the closeness of the actual outcome. That is, participants' cause ratings did not differ between situations in which B just went through the gate (clips 8, 10, 12: $M = .51$, $SD = .40$) compared to situations in which B clearly went through (clips 13 - 18: $M = .49$, $SD = .42$). Similarly, prevention judgments were not different between situations in which B just missed (clips 7, 9, 11: $M = -.41$, $SD = .43$) and situations in which B clearly missed (clips 1-6: $M = -.47$, $SD = .44$).

People's cause and prevention judgments were very well predicted by the PSM. In order to judge whether ball A caused or prevented ball B, participants appear to compare what actually happened with what they think would have happened had A not been present. This very high correlation is achieved without the need for any free parameters in the model. $P(B|A)$ is determined by the outcome of the clip and $P(B|\neg A)$ by participants' judgments in Experiment 2. The PSM also correlates highly with participants' causal attributions when $P(B|\neg A)$ is treated as a free parameter and estimated via the noisy Newtonian physics model with a maximum correlation of $r = .99$ (cf. Figure 2).

The AFM which assumes that people arrive at their judgments via comparing instantaneous force vectors, rather than a mental simulation of the full physical dynamics cannot capture people's judgments equally well. Clip 14 (see Figure 1) gives an example in which the AFM gets it wrong. While participants indicate that A caused B to go through the gate (see Figure 3), the AFM model cannot predict this. In this situation, the angle between the velocity vector of B shortly after the collision and the goal vector α is greater than the angle between the velocity vector of B shortly before the collision and the goal vector β. Hence, the model predicts that A is preventing B but since B does in fact go in, the model's prediction is fixed to 0. In defense of the AFM, it could be argued that clip 14 is better thought of as a causal chain in which A causes B to hit the wall which then causes B to go in. Whether participants would count the static wall as a cause of B going through the gate is an empirical question. In any case, the other problematic clips mentioned above remain. Each of these clips only involves a single interaction.

Experiment 4: Almost Caused/Prevented

The results of Experiment 3 show that people's cause and prevention judgments are only influenced by their degree of belief about whether the event of interest would have happened without the cause being present and not influenced by how close the outcome actually was. However, often the closeness with which something happened clearly matters to us, such as when we almost missed a flight to Japan or only just made it in time for our talk.

As mentioned above, one of the appeals of process accounts is that they acknowledge the semantic richness of the concept of causation by making predictions about which out of several causal verbs people will choose to describe a particular situation. In this experiment, we will demonstrate that our framework is not only capable of capturing the difference between different causal verbs such as *caused* or *helped* but also predicts when people make use of intrinsically counterfactual concepts such as *almost caused* or *almost prevented*. Current process accounts (e.g. Wolff, 2007) cannot make predictions in these situations as they aim to analyze causality without making reference to counterfactuals.

In Experiment 4, participants (N = 41) had to select from a set of seven sentences the one that describes the situation best. The sentences were: A caused / helped / almost caused B to go in the hole; A prevented / helped to prevent / almost prevented B from going in the hole; A had no significant effect on whether B went in the hole or not.

Table 1: Predicted probability of choosing different sentences in Experiment 4.

	outcome	probability
(1) caused	hit	$1-P(B\|\neg A)$
(2) helped	hit	$1-abs(0.5-caused)^*$
(3) almost caused	miss	$p(almost\ B\|A)-prevented$
(4) prevented	miss	$p(B\|\neg A)$
(5) helped to prevent	miss	$1-abs(0.5-prevented)^*$
(6) almost prevented	hit	$p(almost\ \neg B\|A)-caused$
(7) no effect	hit/miss	$1-max((1),...,(6))$

*rescaled to range from 0 to 1

Model Predictions

Table 1 gives an overview of the model predictions which are a function of the outcome, that is, whether B went in or not, and the probabilities $P(B|\neg A)$, $P(almost\ B|A)$ and $P(almost\ \neg B|A)$. For $P(B|\neg A)$ we can again use participants' judgments from Experiment 2 or the predictions of the PSM.

The model's predictions for *caused* and *prevented* are identical to the predictions in Experiment 3. According to our model, the difference between *caused* and *helped* is an epistemic one. People are predicted to select *helped* when B went in and when they were uncertain about what would have happened had A not been present. However, when it was clear that B would have gone in or would have missed, people are predicted to select *no effect* or *caused*, respectively, and not *helped*. Similarly, when B missed and it was uncertain whether B would have gone in, the model predicts that people select *helped to prevent*.

In order to predict when people select *almost caused* or *almost prevented*, we first have to define the probabilities $P(almost\ B|A)$ and $P(almost\ \neg B|A)$. These probabilities express the closeness of an alternative counterfactual outcome to the actual outcome. One way to get at the probabilities would be to have participants judge how closely B hit or missed the gate. However, here we used a variation of the PSM to generate these probabilities. For each clip we ran a set of 100 x 10 noisy simulations for different noise levels from SD = 1° to 5°, whereby the noise was again introduced at the time of collision. If the outcome in the noisy simulations was different from the original outcome in *any* of the ten repetitions in each of the 100 simulated worlds, we counted this as a positive instance. If the outcome in all ten repetitions was the same as the original outcome, we counted this as a negative instance. For example, a value of $P(almost\ \neg B|A) = .87$ in a situation in which B goes through the gate in the original clip, means that in 87 out of the 100 generated worlds, the ball did not go through the gate in at least one of the ten repetitions of each of the worlds. For the remaining 13 worlds, the ball did go in for all ten repetitions. Intuitively, in situations in which the outcome was close, the chances that the outcome in the noisy simulation will be different from the outcome in the original clip in at least one out of ten repetitions are high. However, if ball B clearly missed, for example, it is unlikely that there will be a noisy simulation in which the introduced angular perturbation is sufficient to make B go in.

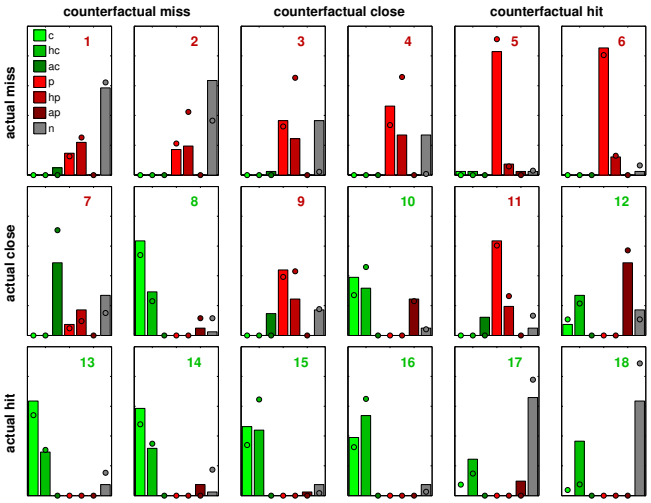

Figure 4: Frequencies with which different sentences were selected in Experiment 4 (bars) and predictions by the Physics Simulation Model (circles), $r = .86$. The color of the clip number indicates whether the ball went in (green) or not (red). *Note*: c = caused, hc = helped (to cause), ac = almost caused, p = prevented, hp = helped to prevent, ap = almost prevented, n = no significant effect.

The model predicts that people will select *almost caused* when B just missed (which means that $P(almost\ B|A)$ is high) and the probability that it would have gone in given that A was absent is low. People should select *almost prevented* when B just went in ($P(almost\ \neg B|A)$ is high), and when it was clear that B would have gone in had A been absent ($P(B|\neg A)$ is high). Finally, if none of these calculations result in a high value, people are predicted to select that A had *no significant effect* on whether B went through the gate.

The model predictions for the 18 different clips can be seen in Figure 4. We used Luce's (1959) choice rule to transform the different probabilities into predictions about the frequencies with which the different sentences will be selected. The model predicts that the modal choice in situations in which B does not go in changes from *prevented* for clips in which it was clear that B would have gone in (clips 5 & 6) to *helped to prevent* in situations in which it was unclear whether B would have gone in (clips 3 & 4). The same switch of the modal response as a function about the certainty of what would have happened is predicted for *caused* (clips 13 & 14) and *helped* (clips 15 & 16). If there was little uncertainty about the counterfactual outcome and it matches the actual outcome, people are predicted to select *had no effect* (clips 1, 2, 17, 18). For clip 7 in which B just misses, people are predicted to select *almost caused* since it was clear that B would have missed but for A. Conversely, for clip 12, the model predicts that people will select *almost prevented*: B just goes in and it was clear that it would have gone in without A being present.

Results and Discussion

The model predicts the frequencies with which participants select the different sentences very well, $r = .86$ (cf. Figure 4). Figure 2 shows the correlation when we generate $P(B|\neg A)$

through noisily perturbing the vector rather than taking participants' ratings from Experiment 2. It predicts the modal choice correctly in 12 out of 18 clips. While participants' modal response does not change between clips 5 & 6 and clips 3 & 4 as predicted by the model, the proportion of *helped to prevent* selections clearly increases. A similar shift is observed between clips 13 & 14 and 15 & 16 for which participants' selection of *helped* increases as a function of the uncertainty over what would have happened.

As predicted by the model, participants' modal response in clip 7 is *almost caused* and in clip 12 *almost prevented*. The variance in responses within a clip is greater for the clips in which the actual outcome was close (middle row) compared to when it clearly missed (top row) or clearly hit (bottom row). For example, in clip 10 in which B just goes in and the counterfactual outcome is close the majority of participants selected *caused* or *helped* while a minority of participants selected *almost prevented*. This pattern closely matches the predictions of our model. Whether a participant is expected to select *caused* or *almost prevented* depends on the participant's subjective belief about the counterfactual outcome. If a participant thought that B would have missed she will say A *caused* or *helped* it. However, if a participant thought that B would have gone in but for A he will select *almost prevented* because B barely goes in.

The close fit between our model prediction and participants' selection of sentences demonstrates that our model is capable of capturing some of the richness of people's causal vocabulary. Our model not only allows to distinguish cases of *causing/preventing* from *helping* but also accurately predicts people's *almost caused/prevented* judgments. Process theories that analyze different causal concepts without the use of counterfactuals (e.g. Wolff, 2007) cannot make predictions about when people will say that something almost happened.

General Discussion

In this paper, we developed a framework for understanding causal attributions that aims to break the longstanding dichotomy between process accounts and dependency accounts of causation. We showed that people's quantitative cause and prevention judgments (Experiment 3) as well as people's use of different causal verbs (Experiment 4) can be very well predicted by assuming that people compare what actually happened when the cause was present, with what they think would have happened in the absence of the cause. We provided evidence that people use their intuitive understanding of physics to simulate possible outcomes (Experiments 1 & 2). Understanding causal attributions in terms of counterfactuals defined over probabilistic generative models sidesteps the presumed trade-off between flexibility and richness described above. Our model retains the generality of dependency accounts while the use of a generative model based on Newtonian physics allows us to capture some of the richness of people's concept of causation.

According to our account, causal attributions are subjective and model-dependent. Two observers with a different understanding of the underlying generative model are predicted to reach different causal verdicts for the same clip when their beliefs about what would have happened in the absence of the cause event differ. The noisy Newtonian physics model predicted participants' judgments well in our experiments. However, we are not committed to this particular generative model – indeed, our account predicts that the ways in which people's intuitive understanding of physics is biased will be mirrored in their causal attributions.

While our framework shares some of the key insights of Wolff's (2007) force dynamic account, such as the need for a richer specification of people's causal representations, our proposals are different in critical respects. Most importantly, our accounts differ in the role that counterfactuals play. Wolff (2007) aims to reduce causal attributions to configurations of force vectors and argues that these force representations (which are primary) can then be used for the simulation of counterfactual outcomes (which are secondary). Our account, in contrast, does not try to explain causal attributions in terms of non-causal representations but postulates that causal attributions are intimately linked with the simulation of counterfactuals. Hence, we claim that in order to say whether A caused B, it is necessary to consider what would have happened to B in the absence of A and not sufficient to only consider what forces were present at the time of interaction between A and B.

In future experiments, we will investigate how our account can handle more complex physical interactions and interactions between intentional agents.

Acknowledgments

We thank Tomer Ullman, Peter Battaglia and Christos Bechlivanidis for very helpful discussions. This work was supported by a doctoral grant from the AXA research fund (TG), a John S. McDonnell Foundation Scholar Award (NG), a ONR grant N00014-09-1-0124 (NG, JT), an ESRC grant RES-062- 33-0004 (DL) and ONR MURI grants W911NF-08-1-0242 and 1015GNA126 (JT).

References

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.
Dowe, P. (2000). *Physical causation*. Cambridge University Press.
Hamrick, J., Battaglia, P., & Tenenbaum, J. (2011). Internal physics models guide probabilistic judgments about object dynamics. In *Proceedings of the 33rd annual conference of the cognitive science society.*
Hume, D. (2000 [1748]). *An enquiry concerning human understanding: A critical edition*. Oxford University Press.
Lewis, D. (1970). *Counterfactuals*. Blackwell.
Luce, R. (1959). *Individual choice behavior: A theoretical analysis*. John Wiley.
Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.
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.
Walsh, C. R., & Sloman, S. A. (2011). The meaning of cause and prevent: The role of causal mechanism. *Mind & Language*, *26*(1), 21–52.
Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.