From counterfactual simulation to causal judgment

Tobias Gerstenberg[1] (tger@mit.edu), Noah D. Goodman[2] (ngoodman@stanford.edu), David A. Lagnado[3] (d.lagnado@ucl.ac.uk) & Joshua B. Tenenbaum[1] (jbt@mit.edu)

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

[2]Department of Psychology, Stanford University, Stanford, CA 94305

[3]Cognitive, Perceptual and Brain Sciences, University College London, London WC1H 0AP

Abstract

In this paper, we demonstrate that people's causal judgments are inextricably linked to counterfactuals. In our experiments, participants judge whether one billiard ball *A* caused another ball *B* to go through a gate. Our *counterfactual simulation model* predicts that people arrive at their causal judgments by comparing what actually happened with the result of mentally simulating what would have happened in the relevant counterfactual world. We test our model against actualist theories of causation which aim to explain causation just in terms of what actually happened. Our experimental stimuli contrast cases in which we hold constant what actually happened but vary the counterfactual outcome. In support of our model, we find that participants' causal judgments differ drastically between such cases. People's cause and prevention judgments increase with their subjective degree of belief that the counterfactual outcome would have been different from what actually happened.

Keywords: causality; attribution; counterfactuals; intuitive physics.

Introduction

In American Football, pass interference is a foul whereby a defender hinders a receiver from catching the ball prior to the ball's arrival. However, a defender's play is only penalized when the referees rule that it would have been possible for the receiver to catch the ball had the defender not interfered. In other words, the defender's action must have made a difference to the outcome to be called a foul.

The pass interference rule embodies a counterfactual criterion of causation. According to a counterfactual theory, one event *C* is a cause of another event *E* if (i) both *C* and *E* actually happened, and (ii) *E* would not have happened if *C* had not happened. While the idea that at least part of what it means to be a cause is to have made a difference to the outcome is intuitively appealing, counterfactual theories have been criticized on various grounds. On the one hand, counterfactuals seem to be doing too much: why consider what would have happened in another possible situation and not just look at what actually happened? On the other hand, counterfactuals seem to be doing too little: counterfactual dependence seems too abstract and indirect. It cannot capture the sense of "oomph" that we get in paradigmatic cases of physical causation (Schaffer, 2005). For example, when we see the collision of two billiard balls, it seems like we can directly perceive causality without considering counterfactuals (Michotte, 1946/1963).

These limitations of the counterfactual framework have motivated a very different way of trying to analyze causation. According to *actualist theories* of causation, causal relationships are determined just in terms of what actually happened. For example, in Dowe's (2000) *conserved quantity theory*, what it means to say that one object caused a change to another object depends critically on the transfer of a physical quantity (such as force or momentum) from the former to the latter. Counterfactual considerations about what would have happened in other possible worlds are irrelevant.

In psychology, the most comprehensive account of an actualist theory of causation has been developed by Wolff (2007). In his *force dynamics model*, causality is reduced to configurations of forces that are present at the time of interaction between an agent and a patient. According to this view, an agent *caused* a patient to reach a certain endstate if the following conditions are met: (i) the patient did not have a tendency toward the endstate and (ii) the agent and patient force vectors combined to a resultant force pointing toward the endstate.[1]

In our work, we combine what we see as the strengths of the counterfactual and actualist accounts of causation. Our *counterfactual simulation model* maintains the view that people's causal attributions are intrinsically connected to whether the event of interest made a difference to the outcome. This aspect of causation is well-captured by counterfactual theories. However, in line with Wolff's (2007) *force dynamics model*, we acknowledge that people's intuitive theories are often much richer than what can be expressed in terms of the formal accounts of counterfactual reasoning to date (e.g. Halpern & Pearl, 2005). We assume that people use their intuitive domain theories to simulate what would have happened in the relevant counterfactual world (cf. Battaglia, Hamrick, & Tenenbaum, 2013). Consider stepping into the shoes of a football referee to decide whether a given football was catchable. Being able to mentally simulate the counterfactual requires both a sophisticated understanding of how people work (e.g., did the receiver try to catch the ball?) as well as how the world works (e.g., was the distance between the ball and the receiver such that it would have been physically possible for him to catch the ball?).

There is a rich literature on causal attribution which discusses the tension between counterfactual (e.g. Kahneman & Tversky, 1982) and actualist determinants (e.g. Mandel, 2003) of people's causal judgments. Studies in this literature

[1]Because the *force dynamics model* reduces causation to configurations of forces, a patient's tendency is not defined counterfactually in terms of whether or not the patient would have reached the endstate in the absence of the agent. Rather, tendency is defined as the direction in which the patient's force points at the time of interaction between agent and patient. More recently, the counterfactual concept of a *virtual force* has been incorporated into the *force dynamics model* to explain people's causal judgments for situations that involve omissions and/or chains of events (see Wolff, Barbey, & Hausknecht, 2010).

usually ask participants to reach their causal verdicts based on written vignettes which stipulate explicitly what would have happened in the relevant counterfactual worlds. In our work, in contrast, we present participants with animated clips of causal interactions in a physical domain. This allows us to manipulate people's uncertainty in the relevant counterfactual outcome in a quantitative way and thus test more rigorously the relationship between counterfactual and causal judgments.

In previous work (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012), we provided evidence that the *counterfactual simulation model* accurately predicts people's causal judgments. However, these experiments were not designed to test the broader claim that causal judgments are intrinsically linked to counterfactuals and that actualist theories cannot in principle explain people's judgments.

In this paper, we provide a stronger test of the role that counterfactuals play for causal attributions. We contrast pairs of situations in which we match exactly what actually happened but vary what would have happened in the relevant counterfactual world. Furthermore, we test situations in which the counterfactual outcome is held constant but the actual outcome is brought about in different ways. The results support the *counterfactual simulation model* over actualist theories of causation: while participants' causal judgments are strongly affected by manipulating the relevant counterfactual, their judgments are much less influenced by the way in which the actual outcome was brought about.

Counterfactual Simulation Model

The *counterfactual simulation model* of causal attribution applies to any domain where people are able to simulate what would have happened in the relevant counterfactual world. Here, we will focus on judgments about the consequences of collisions between two billiard balls. The causal question of interest is whether ball A caused ball B to go through a gate (or prevented it from going through). Figure 1 shows four diagrammatic illustrations of clips used in the experiment. The solid lines show the actual paths of ball A's and B's movement up until the point of collision and the path that ball B actually traveled after the collision. The dashed lines show the path that B would have traveled if ball A had been removed from the scene.

We will now discuss how our general framework predicts people's causal and counterfactual judgments.

Causal judgments

Our *counterfactual simulation model* predicts that people's causal judgments are a function of their subjective degree of belief that the causal event made a difference to whether or not the outcome would occur. People are predicted to compare the actual outcome with their belief about what the counterfactual outcome would have been. We capture the extent to which a cause is believed to have made a difference to the outcome in terms of Pearl's (1999) counterfactual definition of the probability that the cause event X was necessary for the

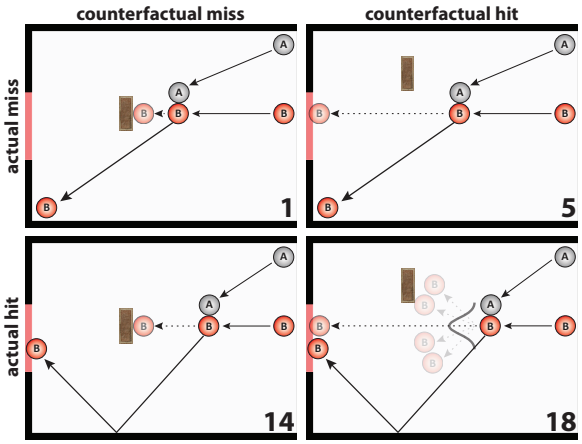


Figure 1: Diagrammatic illustrations of four stimuli. In each pair of stimuli (top vs. bottom pair), what actually happened is held constant. The location of the brick only influences the counterfactual outcome. *Note*: Clip 18 illustrates the *noisy physics model*.

effect event Y to occur. In our case, X denotes the event that there was a collision between A and B which either occurred (x) or didn't occur ($\neg x$). Y denotes the event that B went through the gate which, again, either occurred (y) or didn't occur ($\neg y$).

The probability PN_{caused} that the collision event between A and B was a necessary cause of B going through the gate is defined in the following way:

$$PN_{caused} = P(\neg y_{do(\neg x)} | x, y) \quad (1)$$

In words, PN_{caused} denotes the probability of the counterfactual that ball B would *not* have gone through the gate, if there had been *no* collision between A and B ($\neg y_{do(\neg x)}$), given that in the actual world, A and B collided (x) and ball B went through the gate (y).

Similarly, we can define the probability $PN_{prevented}$ that the collision between A and B was a necessary cause of B *not* going through the gate as:

$$PN_{prevented} = P(y_{do(\neg x)} | x, \neg y) \quad (2)$$

In words, $PN_{prevented}$ denotes the probability of the counterfactual that ball B would have gone through the gate in the absence of the collision ($y_{do(\neg x)}$), assuming that in the actual world, the balls collided (x) and B did in fact *not* go in ($\neg y$).

Our *counterfactual simulation model* shares with Pearl (1999) the idea that people compute the relevant counterfactuals by manipulating an intuitive causal domain theory. However, our model differs from Pearl's account in both the form of the causal model and the nature of the manipulation. First, while Pearl represents causal knowledge in terms of graphs and structural equations, we assume a richer dynamic representation akin to a physics engine (cf. Battaglia et al., 2013). Second, whereas Pearl defines the counterfactual manipulation in terms of a formal operation on a graph, in our model, we compute the probability of counterfactuals via removing ball A from the world shortly before A and B would have collided and then simulating what would have happened.

We predict that people's cause and prevention judgments increase with PN_{caused} and $PN_{prevented}$, respectively. People are predicted to say that ball *A* caused ball *B* to go through the gate when *B* did in fact go in, and when they are confident that *B* would not have gone in had the collision not taken place. Similarly, we predict that people give high prevention ratings when *B* missed the gate and they are confident that it would have gone in without the collision. We predict intermediate cause and prevention judgments when people are unsure about what the counterfactual outcome would have been (i.e. for intermediate values of PN_{caused} and $PN_{prevented}$). Finally, we predict that people will say that *A* neither caused nor prevented *B* from going through the gate when they are confident that the counterfactual outcome would have been the same as the actual outcome (i.e. when PN_{caused} or $PN_{prevented}$ is low).

Counterfactual judgments

In our experiments, we ask people to evaluate the counterfactual of whether ball *B* would have gone through the gate if ball *A* had not been present in the scene. We assume that people arrive at their belief about what would have happened in the relevant counterfactual world by using their intuitive understanding of the domain. In our case, the domain of interest comprises basic physical concepts such as velocity and force as well as non-physical concepts such as teleportation.

Previous work has shown that people's predictions in certain physical domains are well-explained as an approximation to Newtonian physics (Battaglia et al., 2013; Sanborn, Mansinghka, & Griffiths, 2013). In line with this work, we capture people's uncertainty in their mental simulation of the counterfactual by introducing Gaussian noise to ball *B*'s velocity vector at the time at which the collision between *A* and *B* would have taken place (see Figure 1, clip 18). It is at this point in time that the relevant counterfactual world diverges from the actual world and participants are required to rely on their mental simulation to predict whether or not ball *B* would have gone in. By adding Gaussian noise to *B*'s velocity vector, we capture the dynamic uncertainty inherent in people's intuitive physical model (cf. Smith & Vul, 2012). In order to predict people's counterfactual judgments, we first run a large number of simulations for each clip in which we randomly perturb *B*'s velocity vector as described above. We then compare the proportion of times in which ball *B* went through the gate in the sample of simulations to people's judgments. For example, if ball *B* went through the gate in all of the generated samples, we predict that people should be very certain that *B* would have gone in. We predict that participants should be maximally uncertain about the counterfactual outcome when the generated sample is exactly split between clips in which *B* went in and didn't go in.

Experiment

In all of the clips used in the experiment, both balls enter the scene from the right and collide once with each other. In addition to the solid walls that mark the border of the scene, some clips also featured a brick and/or a teleport.[2]

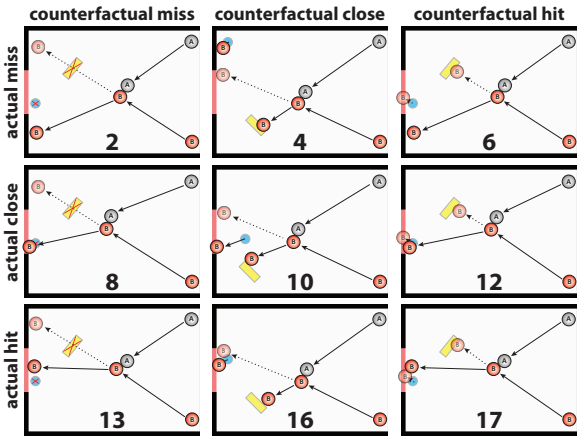


Figure 2: Diagrammatic illustrations of nine stimuli used in the experiment. *Note*: If the teleport is active, ball *B* is teleported from the yellow rectangle to the blue circle (see, e.g. clip 10).

While the location of the solid walls was fixed, the placement of the brick and teleport changed between the clips. Participants were instructed that the brick was solid and that the balls bounce off it in case of collision. Participants were also instructed about how the teleport worked. The teleport only affected ball *B* but not ball *A*. Ball *B* always exited the teleport through the blue circle in the same direction in which it had entered through the yellow rectangle (see, e.g., Figure 2, clip 10). Participants also learned that the teleport was sometimes deactivated which was marked by a red cross on top of the teleport's entrance and exit. If the teleport was deactivated, the teleport's entrance and exit had no influence on *B*'s movement.

The stimuli were designed with two goals in mind: First, in order to test the *qualitative hypothesis* that causal judgments are intrinsically linked to counterfactual considerations, we created pairs of stimuli that held constant what actually happened and only varied what would have happened in the relevant counterfactual world. Figure 1 shows two pairs of clips in which differences in the relevant counterfactual worlds were achieved by varying the placement of a brick. For example, in both clips 1 and 5, the interaction between *A* and *B* is exactly the same (as illustrated by the identical solid lines in both diagrams). However, whereas in clip 1, ball *B* would have been blocked by the brick if ball *A* had not been present, in clip 5, *B* would have gone through the gate in *A*'s absence. Similarly, in both clips 14 and 18, *A* collides with *B* and *B* goes through the gate after having bounced off the wall. While in clip 14, *B* would have been blocked by the brick if *A* hadn't been present, in clip 18, *B* would have gone through the gate even in the absence of *A*.

While we had to vary the placement of the brick in order to generate the desired contrast between counterfactual worlds, the inclusion of the teleport allowed us to achieve the same effect without having to change the placement of any compo-

[2]The different clips as well as a demo of the experiment may be accessed here:
`http://web.mit.edu/tger/www/demos/teleport_demos.html`

nents. We simply contrasted cases in which the teleport was either on or off. For example, while in both clips 13 and 17, the collision event between A and B is identical, B would have not gone through the gate in the absence of A when the teleport was off (clip 13), whereas it would have gone through when the teleport was on (clip 17). The teleport further allowed us to test the flexibility of people's counterfactual simulations.

Second, in order to test the *quantitative hypothesis* about the relationship between people's causal judgments and their subjective degree of belief about the counterfactual outcome, we crossed how closely ball B actually went through the gate with how closely B would have gone through the gate if ball A had not been present in the scene (see Figure 2). For example, in clip 6, ball B actually missed the gate. However, if A had not been present, B would have gone through the gate via the teleport. We generated two clips for each of the nine cells that cross the closeness of the actual and counterfactual outcome.

Methods

Participants and materials 80 participants (43 female, $M_{age} = 33.94$, $SD_{age} = 12.24$) were recruited via Amazon Mechanical Turk and paid $1.5 compensation. The experiment was programmed in Flash CS5 and the clips were generated using the physics engine Box2D.

Design and procedure Each participant saw two blocks of trials: in the *counterfactual block*, the clips were paused at the time of collision. After having seen the clip twice, participants were asked to judge whether ball B would have gone through the gate if ball A had not been present in the scene. Participants indicated their response on a slider whose endpoints and midpoint were labeled "definitely no" (0),"definitely yes" (100) and "unsure" (50), respectively. After having indicated their response, participants received feedback by being shown the full clip in which ball A was removed from the scene.

In the *causal block*, participants saw each clip played twice until the end and then asked: "What role did ball A play?" The endpoints were labeled "it prevented ball B from going through the gate" (-100) and "it caused ball B to go through the gate" (100). The midpoint was labeled "neither" (0). Participants were instructed to use intermediate values to indicate that A somewhat caused B to go through the gate (or somewhat prevented it from going through).

We counterbalanced the order of the two types of blocks. Half the participants made counterfactual judgments before causal judgments and vice versa. Each block started with two practice clips. The 18 test clips were randomized within blocks. We also counterbalanced the vertical position of the balls and the other components. On average, the experiment took 21.16 minutes ($SD = 5.08$) to complete.

Results and Discussion

Counterfactual judgments Since there was neither a significant main effect of block order (i.e. whether participants answered the counterfactual or causal questions first), $F(1,78) = 0.59, p = .45$, nor an interaction effect of block order and clip number, $F(17,1326) = 0.31, p = .99$, we aggregated the counterfactual judgments across both groups of participants. In order to model participants' counterfactual judgments, we created different noisy physics models by varying the degree of Gaussian noise (from $SD = 1^\circ$ to 20°) that was applied to perturb B's velocity vector in the counterfactual world (see Figure 1, clip 18).

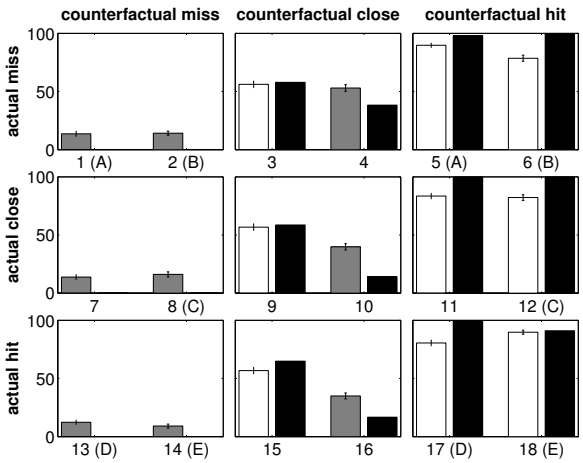
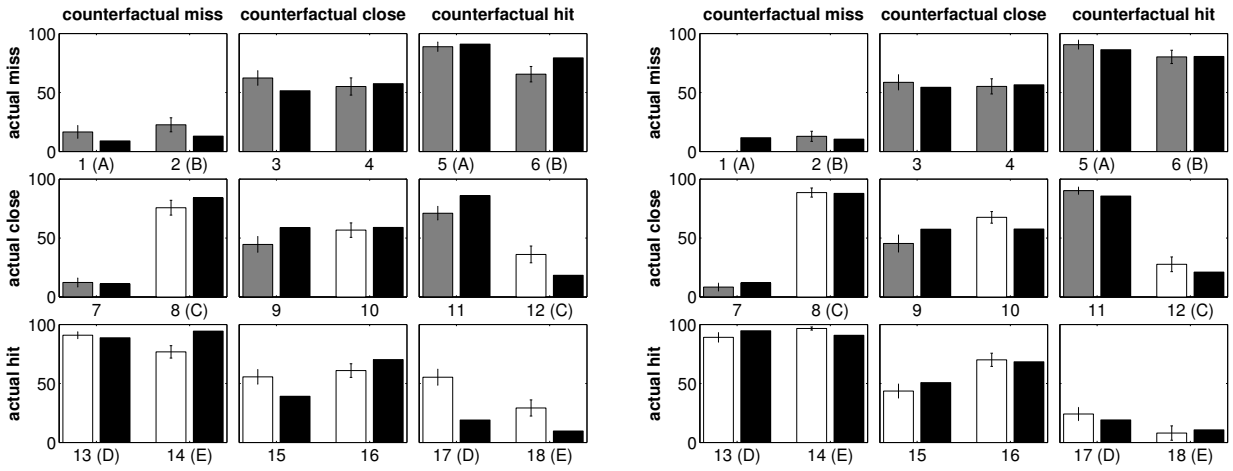


Figure 3: Mean counterfactual judgments for situations in which B would have missed (gray) or gone in (white) together with the predictions of the noisy physics model (black bars). Labels (A) – (E) indicate pairs of clips which matched what actually happened (e.g. clips 1 and 5 in Figure 1). Error bars in all figures indicate *SEM*.

Figure 3 shows participants' mean counterfactual judgments together with the predictions of the best-fitting noisy simulation model ($r = .98, RMSE = 14.64$ for $SD = 4^\circ$). The model accurately captures that participants are certain about the counterfactual outcome for those cases in which ball B would clearly have missed (left column) or clearly gone through the gate (right column). It also captures participants' uncertainty for those cases in which the counterfactual outcome would have been close (middle column). For comparison, the correlation between participants' counterfactual judgments and the simulation model for $SD = 0^\circ$ (i.e. no noise), 10°, and 20° was $r = .89$, .89 and .65, respectively. A model without noise fails to capture people's uncertainty in those cases in which the counterfactual outcome was close.

The high correlation between the noisy physics model and participants' counterfactual judgments demonstrates that within our domain, people's counterfactual predictions are well approximated by a noisy version of Newtonian physics. The finding that people had no trouble in predicting the counterfactual outcome in situations that involved the teleport demonstrates the flexibility of people's mental simulations (see Figure 2, clips 6, 12, and 17).

Causal judgments There was a significant interaction effect of block order and clip number on participants' causal judgments, $F(17,1326) = 3.29, p < .001$. However, since the qualitative pattern of results was very similar between



(a) Causal before counterfactual judgments

(b) Counterfactual before causal judgments

Figure 4: Mean cause (white) and prevention (gray) judgments together with the predictions of the *counterfactual simulation model* (black bars) separated according to the block presentation order. Labels (A) – (E) indicate pairs of clips which matched what actually happened.

the two conditions, we will first focus on the commonalities between the conditions before discussing the differences between them.

The *counterfactual simulation model* predicts a close coupling between people's counterfactual and causal judgments. Cause and prevention judgments are predicted to increase with people's beliefs that the counterfactual outcome would have been different from the actual outcome. Figure 4 shows participants' mean cause and prevention judgments together with the predictions of the *counterfactual simulation model* separated by block order. We reverse coded the prevention ratings so that both cause and prevention ratings are on a scale from 0 to 100. As predicted, participants' causal judgments differed significantly between situations in which we held constant what actually happened but varied what would have happened in the counterfactual world via changing the placement of the brick (Figure 1) or turning the teleport on or off (Figure 2). Across both conditions, prevention judgments were significantly higher for clips in which ball B would have gone in compared to clips in which B would have missed (see pairs of clips with the labels A and B in Figures 4a and 4b), $t(79) = 16.87, p < .001, d = 1.89$. Cause judgments were significantly higher for the clips in which B would have missed compared to the matched clips in which it would have gone in (see pairs C–E), $t(79) = 12.87, p < .001, d = 1.44$.

We used participants' own counterfactual judgments to determine the values of PN_{caused} and $PN_{prevented}$ for the different clips. The *counterfactual simulation model* explains participants' causal judgments very accurately, $r = .96, RMSE = 8.33$. The model achieves this high explanatory fit without the need for any free parameters. The correlation between the model and participants' causal judgments is similarly high if we determine PN_{caused} and $PN_{prevented}$ based on the best-fitting noisy physics simulation model with $SD = 4°$, $r = .93, RMSE = 18.95$.

Our model accurately predicts that participants' *prevention* judgments increase from situations in which the counterfactual outcome was a clear miss (clips 1, 2, and 7), close (clips 3, 4, and 9), to where ball B would clearly have gone in (clips 5, 6, and 11), $F(2,158) = 160.4, p < .01$. Conversely, participants' *cause* judgments decreased from situations in which B would have clearly missed (clips 8, 13, and 14) to where it would have been close (clips 10, 15, and 16) to where B would clearly have gone in (clips 12, 17, and 18), $F(2,158) = 93.05, p < .01$.

Taken together, these results support both our qualitative and quantitative hypotheses about the relationship between counterfactuals and causal judgments. Causal judgments differed significantly between clips in which what actually happened was held constant. Furthermore, causal judgments were closely linked to people's uncertainty in the counterfactual outcome. Cause and prevention ratings were low when the counterfactual outcome was believed to be identical to the actual outcome (e.g., clips 1, 2 and 17, 18). Participants' gave intermediate judgments when they were unsure about what the counterfactual outcome would have been (e.g., clips 3, 4 and 15, 16). Judgments were highest when participants were confident that the counterfactual outcome would have been different from the actual one (e.g., clips 5, 6 and 13, 14). Finally, the results also show that differences in how the actual outcome came about had almost no effect on participants' judgments. This can be seen by comparing cases in which the closeness of the actual outcome was varied but the closeness of the counterfactual outcome held constant (cf. clips with the same outcome between different rows in Figure 4, such as clips 13 and 14 vs. clip 8).

Order effects As can be seen by contrasting Figures 4a and 4b, the overall differences between the two block orders were small with a mean absolute difference between clips of 10.47 ($SD = 7.65$). The *counterfactual simulation model* predicted participants' causal judgments better when they answered the counterfactual questions first ($r = 0.98, RMSE = 6.03$) compared to when the order of judgments was reversed ($r = 0.91, RMSE = 14.26$). In the causal-judgments-first condition, the correlation between counterfactual and causal judgments was significant on the individual participant level for

22 out of the 40 participants with a median correlation over all 40 participants of $r = .49$. In the counterfactuals-first condition, the correlation was significant for 29 out of 40 participants with a median correlation of $r = .64$.

The differences between the two order conditions were strongest for the situations in which *B* went through the gate and it was clear that it would have gone through even in the absence of *A* (i.e. clips 12, 17, and 18). Participants' causal judgments who experienced the causal block first were significantly higher ($M = 44, SD = 33.56$) compared to those who answered the counterfactual questions first ($M = 23.86, SD = 30.71$), $t(78) = 2.8, p < .01, d = 0.57$. Overall, these order effects suggest that there was a stronger influence of counterfactual judgments on causal judgments than vice versa.

General Discussion

The results of our experiment demonstrate that people's causal judgments are inextricably linked to counterfactuals. As predicted by our *counterfactual simulation model*, people make causal judgments by comparing what actually happened with what they think would have happened in the counterfactual world in which the causal event of interest hadn't taken place. They use their intuitive understanding of the domain in order to simulate what would have happened in the relevant counterfactual world. These counterfactual simulations are not limited to physical interactions but may also include more abstract interactions such as teleportation.

We capture people's uncertainty in the counterfactual outcome by assuming that their mental simulations of what would have happened are somewhat noisy (cf. Smith & Vul, 2012). The *counterfactual simulation model* accurately predicts the close relationship between people's uncertainty in the counterfactual outcome and their causal judgments. People's cause and prevention judgments increase with their subjective degree of belief that the causal event of interest made a difference to the outcome.

By contrasting situations in which we held constant what actually happened and only varied what would have happened in the relevant counterfactual world, our experiments constitute the strongest possible test between actualist and counterfactual theories of causal judgment. The fact that participants' judgments differed dramatically between these cases provides strong evidence for our *counterfactual simulation model* and against the possibility of giving an adequate account of people's causal judgments within an actualist framework.

Most participants' causal judgments were highly correlated with their own counterfactual judgments and the correlation was particularly strong for the group of participants who was asked to make counterfactual judgments first. We take this as evidence that while most people seem to naturally consider counterfactuals when making their causal judgments, some participants might have been prompted to do so via having been asked explicitly.

Finally, let us discuss one limitation of our current model. We explain causal judgments in terms of the subjective degree of belief that the event of interest made a difference to the outcome. However, there are some situations in which participants are sure that the collision between *A* and *B* made no difference as to whether or not *B* would go through the gate but they still give a relatively high causal rating. In clip 17, for example, participants are certain that ball *B* would have gone through the gate via the teleport if there had been no collision with *A* (Figure 3). Nevertheless, participants still say that *A* somewhat caused *B* to go through the gate. This is particularly the case for participants who made causal judgments first (cf. Figures 4a and 4b).

One way to capture what's going on here is to say that people not only care about whether the cause was *necessary* to bring about the outcome in the given situation but also whether it was *sufficient*. In clip 17, the collision event was sufficient because it would have caused *B* to go through the gate even if the teleport had been off. Alternatively, people might not only consider *whether* an event of interest made a difference to the outcome but also care about *how* it did so. In future work, we will explore ways of incorporating the notion of sufficiency into our *counterfactual simulation model*.

Acknowledgments

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

References

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.

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

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

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.

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.

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.

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

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

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.

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

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

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.