Causal Reasoning Across Agents and Objects

Bryan S. Gonzalez (Bryan.S.Gonzalez.GR@Dartmouth.edu)
Department of Cognitive Science, Dartmouth College 5 Maynard Street Hanover, NH 03755 USA

Tobias Gerstenberg (Gerstenberg@Stanford.edu)
Department of Psychology, Stanford University, 450 Jane Stanford Way, Building 420 Stanford, CA 94305 USA

Jonathan S. Phillips (Jonathan.S.Phillips @Dartmouth.edu)
Department of Cognitive Science, Dartmouth College, 5 Maynard Street, Hanover, NH 03755 USA

Abstract

This work attempts to bridge the divide between accounts of causal reasoning with respect to agents and objects. We begin by examining the influence of animacy. In a collision-based context, we vary the animacy status of an object using 3D animations. By holding the fine-grained kinematics of the actual and counterfactual outcomes fixed across animate and inanimate conditions, we find that animacy itself has no effect on causal attribution judgments. Next, we test if causal judgments for animate and inanimate objects differ as a function of the counterfactuals they respectively afford in a disjunctive causal structure. Here, we find that the effect of perceived animacy on causal attribution is mediated by differences in counterfactual judgments. Finally, we introduce the known effect of prescriptive norm violations to this paradigm. Our results collectively highlight how normative expectations specify the counterfactual considerations that guide causal reasoning about both agents and objects.

Keywords: causal reasoning; counterfactuals; animacy; intuitive physics

Introduction

Mechanisms of counterfactual thinking have emerged as a compelling account to explain how humans make causal judgments about physical events. However, it is unclear whether the same processes used to make causal judgments for purely physical events are generalized to situations involving goal-directed agents. Do people use the same mental operations when deciding that a falling tree caused damage to the car as when deciding that a CEO's poor decisions caused the company's bankruptcy?

There is a long history of work on how humans reason about outcomes caused by physical objects (Michotte, 1946). Humans possess internal models capable of simulating the mechanics of rigid bodies in space (Ullman et al., 2017). When considering what caused an outcome event, this intuitive representation of physics is used to constrain the list of possible candidates to only those objects that *could* be causal in accordance with the laws of physics. This mechanistic understanding has aided cognitive scientists studying causal reasoning in purely physical contexts (Gerstenberg et al., 2021), but it remains unknown whether the same process is applicable to situations involving *intentional* agents who cause events.

A relatively separate approach to studying causal cognition has focused on more complicated cases and includes agents who make decisions, as is often of interest in many real-world situations. These research designs typically include verbal descriptions of the events and ask participants for explicit judgments of causality (Alicke, 1992; Knobe & Fraser, 2008; Samland & Waldmann, 2016). A strength of these approaches comes from their ecological validity. However, the qualitative nature of these methods are not well suited to characterizing the cognitive processes involved more mechanistically. This is because, unlike rigid body physics, the factors influencing human behavior are far too vast and, currently, mysterious to be programmed or learned with much fidelity by machines (Bishop, 2020; Fjelland, 2020).

The following studies highlight the role of counterfactual considerations to bridge these separate literatures and support a more unified view of causal cognition. We do this by exploring causal judgments in experiments that vary the agentic status of a candidate cause while keeping other physical dynamics in accordance with the laws of physics. This approach allows us to explore if differences in causal attributions can be explained as a function of the different counterfactuals afforded to agents and objects, or if causal judgments across these domains result from largely distinct cognitive mechanisms.

Experiment 1 manipulates the animacy of a candidate cause as either animate (following a manually specified trajectory simulating the movement of a goal-directed agent), or inanimate (following a trajectory prescribed for objects by physics). By holding the fine-grained kinematics of the actual and counterfactual outcomes fixed across these conditions, we find that perceived animacy has no isolated effect of on causal attribution judgments. Experiment 2 explores whether judgments of agentic and non-agentic causes come apart when the relevance of counterfactuals in overdetermined causal structures is varied. Here, we find that agents can elicit different counterfactuals from objects, resulting in different causal attributions to agents and objects for the same events. Experiment 3 extends this paradigm to focus specifically on the role of intentional or unintentional prescriptive norm violations, as observed in prior work (Henne et al., 2021; Kirfel & Lagnado, 2018; Knobe, 2009). We find that normative expectations dictate the counterfactuals considered when anthropomorphized agents intentionally

violate norms to cause an outcome. These expectations do not extend to inanimate causes. Taken together, this work provides a first step in narrowing the divide between our understanding of the causal attribution processes for animate agents and inanimate objects.

Experiment 1: Manipulating animacy

Prior work provides some evidence that causal cognition may operate differently for animate agents and inanimate objects. For example, people who cause outcomes deliberately, are seen as more causal than people who do so unintentionally (Hilton et al., 2016; Malle et al., 2014), suggesting that causal cognition with respect to agents is sensitive to information about intentionality. Since inanimate objects inherently lack intentionality, people may plausibly use different criteria for judging causation. Alternatively, causal judgments for goal-directed agents and inanimate objects may both rely on a single mechanism that involves evaluations of whether or not the outcome would have obtained in counterfactuals in which the candidate had been altered or removed (Gerstenberg et al., 2021; Kominsky & Phillips, 2019).

The current experiment investigates whether animate agents are judged as causes to the same extent as inanimate objects for the same outcome. Crucially, we isolate the influence of animacy on causal judgments by holding the physical parameters of the causal events and the actual outcomes fixed across animacy conditions. If causal cognition operates differently for goal-directed agents versus inanimate objects, this may be reflected in differences in causal judgments despite these similarities. Alternatively, if the cognitive mechanisms underlying these judgments are instead isomorphic, animated agents and inanimate objects may be judged as equally causal for the same outcomes.

Methods

Participants 105 adults (34 female, 71 male) were recruited from Prolific. All participants were at least 18 years old, (M_{age} = 25, *SD* = 8), endorsed fluency in English, and successfully completed more than 94% of their previous tasks on the recruitment platform.

Stimuli and Procedure Classic research on causal cognition involved the perception of "launching" events in which one object appears to collide with another static object, thereby "causing" its subsequent motion (Michotte, 1946). Following this work, and in each of the experiments we report, the object that *launches* is referred to as the *agent*, whereas the object that is *being launched* is referred to as the *patient*. We refer to "agents" in this sense of launching patients, and as either animate or inanimate.

The video stimuli depicted a billiard table in which an agent ball collides with a moving patient ball, resulting in the patient landing in or missing one of two corner pockets. The patient's trajectory is such that it would land in or miss a corner pocket in a counterfactual scenario in which the agent ball had not been there. Thus, we used a 2 (animacy of agent: animate vs inanimate) x 2 (patient outcome: in vs out of corner pocket) x 2 (patient counterfactual: in vs out of corner pocket) design. The patient ball was always objectively inanimate and moved according to a physics simulation. 24 unique patient trajectories were presented, each with an animate and inanimate agent. We included four (two animate, two inanimate) catch trials, in which the agent does not collide with the patient. All stimulus videos in this and subsequent experiments were created using Blender 3D computer graphics software v2.9, and the Bullet physics engine to simulate ball trajectories and collisions. Crucially, the trajectory and fine-grained kinematics of the patient ball were matched across animacy conditions. In videos containing animate agents, agent ball trajectories before the collision were manually specified along a bezier curve roughly simulating the movement of an animate, goal-directed, agent and culminating at the location, angle, frame number, and with an instantaneous velocity equivalent to that of the inanimate agent in the complimentary stimulus depicting the same patient trajectory (*Figure 1*).

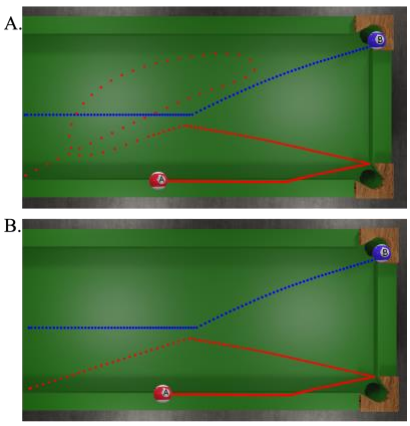

Figure 1: **Experiment 1**. Example trajectores for an agent (red) and patient (blue), the patient outcome changes from a miss, in a counterfactual without the red agent, to actually landing in. A. Animate Agent condition; B. Inanimate Agent condition.

On each trial, participants used a slider bar to rate the extent to which they agreed with a prompt shown below the video. Depending on the outcome of the patient ball, the prompt asked participants to indicate their agreement with the statement "*Ball A* [agent] *caused Ball B* [patient] *to land in (miss) the pocket*". Ratings were made on an integer scale from 0 ("Disagree") to 100 ("Agree"). After indicating their agreement with the causal statement, participants indicated the extent to which they perceived the agent ball ("Ball A") as animate as a manipulation check. All participants observed and made causal and animacy judgments for all 48 stimulus clips. The materials, data, and analysis code for all experiments reported in this paper are available here: https://github.com/PhilLaboratory/CausalAgents.git

Results

Data from 8 participants were excluded from analyses for reporting an agreement rating of ≥ 60 with the causal

statement on catch trials where the agent and patient did not make contact. These catch trials were not further analyzed.

First, we found that our manipulation of animacy was successful. A likelihood ratio test revealed a large difference in animacy rating between trials in which agents' movements were manually manipulated to appear animate (M = 87, SD = 27), and trials in which the agent's movement was rendered directly from a physics simulation (M = 34, SD = 38), $\chi^2(1)$ = 95.11, $p < .0001$.

Using a comparison of linear mixed-effects models, we next tested for a three-way interaction effect between patient outcome × patient counterfactual × agent animacy on causal judgments, including random effects for subject and patient trajectory. This was not significant, $\chi^2(1) = .742$, $p = .389$. We next tested for the two-way interaction effects of patient outcome × patient counterfactual outcome on causal ratings for the agent across animacy conditions. Consistent with existing theories (Woodward, 2003), we found the interaction effect of outcome and counterfactual significantly influenced causal ratings for the agent, $\chi^2(2) = 8.191$, $p = .004$. Planned pairwise comparisons of the four possible outcome/ counterfactual combinations were carried out using the Estimated Marginal Means package in R (Lenth et al., 2022). These tests reveal that participants rated both animate and inanimate agents as more causal when it changes the patient trajectory from (counterfactually) missing to actually landing in, compared to when the patient outcome is unchanged by the agent and misses, $t(27.28) = 2.94$, $p < .05$, or lands in, although this latter difference did not reach significance, $t(26.81) = -2.43$, $p = .09$ (*Figure 2*). In other words, participants rated agents that made a difference to the patient's outcome as more causal, regardless of whether the agent was animate or inanimate. This effect was greater when patients ultimately landed in the corner pocket, than when patients ultimately missed. Crucially, we found no effect of agent animacy on causal ratings, $\chi^2(1) = 0.023$, $p = .88$ (*Figure 2*, blue vs. orange).

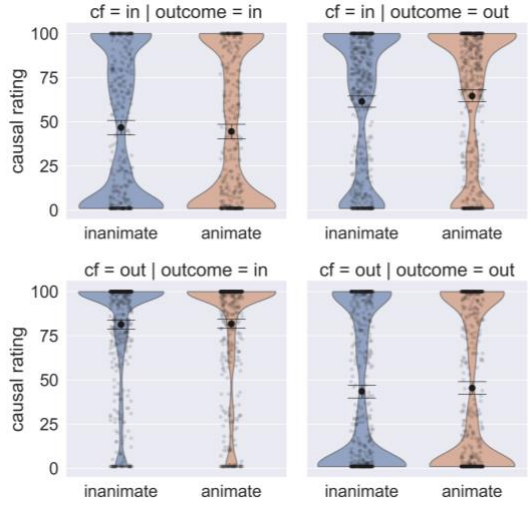

Figure 2: **Experiment 1**. Causal attribution ratings to inanimate and animate agent balls for the patient outcome in all actual - counterfactual outcome pairs. Large points are means with 95% bootstrapped confidence intervals.

Discussion

This experiment provided a simple test of whether causal cognition for intentional agents and inanimate objects operate under unified or distinct mechanisms. In a collision context, we manipulated the animacy of the agent ball, as well as the patient ball's actual and counterfactual outcomes. We found that the interaction of the patient's actual and counterfactual outcome had a significant effect on causal ratings for the agent ball. This interaction was driven by the condition in which the patient's outcome was changed from counterfactually missing the pocket to landing in post-collision. This result may have occurred because there are far fewer ways of hitting the patient into the corner pocket than diverting it away from the corner pocket, making this event relatively less likely. Higher causal judgments in these cases are consistent with existing work demonstrating an inflation in causal judgments for more abnormal events (Hilton & Slugoski, 1986; Icard et al., 2017).

The main variable under investigation in this experiment was the effect of animacy. Despite our clear ability to manipulate perceived animacy, we did not find that this had an influence on causal attribution ratings. This was likely because counterfactual outcomes were precisely matched across the animacy conditions in order to isolate the possible effect of animacy. Our results are consistent with previous studies demonstrating that causal cognition for intentional agents and inanimate artifacts may be underpinned by the same counterfactual mechanisms (Kominsky & Phillips, 2019). We extend those findings here and provide an even more compelling case that includes objects animated to appear intentional.

Experiment 2: Counterfactual relevance

In Experiment 1, we found that animacy, in isolation, does not make a candidate more causal of an outcome, at least when the patient counterfactuals were matched across animate and inanimate conditions. In Experiment 2, we sought to indirectly manipulate the contrast between actual and counterfactual outcomes through the perception of animacy. Keeping outcomes matched across conditions, we examine if animacy affects causal reasoning when it affects the *availability* of difference-making counterfactuals. Specifically, we focus on a case of overdetermination in which the trajectory of both an agent ball and a patient ball are each individually sufficient to bring about the outcome (a tower of blocks falling over). In all cases, the agent ball hits the patient ball, and the patient ball hits a block tower, which falls over. In this causal structure, the possible counterfactual trajectories differ for animate and inanimate agents. Specifically, because the animate agent's actions are self-propelled and goal-directed, there are relevant counterfactuals in which they make a difference in the outcome by prevent the patient ball from colliding with the tower; such counterfactuals are not applicable to inanimate agents. Thus, we expect that any effect of perceived animacy on causal judgment will be mediated by the difference in counterfactual outcomes considered for the animate vs.

inanimate agents. More specifically, the inanimate agent may be judged to be less causal for the same outcome than an animate agent who is perceived as capable of changing the outcome if they wanted.

Methods

Participants 210 adults (M_{age} = 38, *SD* = 14, 107 female, 103 male) were recruited through Prolific. All participants were at least 18 years old, endorsed fluency in English, and successfully completed more than 94% of their tasks in the past on the recruitment platform.

Stimuli and Procedure To manipulate the perception of animacy, a priming clip was shown to participants which involved two balls moving around a platform with a tower of stacked blocks. For participants randomly assigned to the Inanimate Agent condition, the balls were shown colliding with each other and bouncing off different edges of the platform (*Figure 3A*). Movements of both balls in the Inanimate Agent condition and the patient ball in the Animate Agent condition were simulated using the Bullet physics simulation engine. For participants randomly assigned to the Animate Agent condition, the priming clip depicted an inanimate patient ball and an animate agent ball that appeared to be "playing" with the inanimate patient, repeatedly knocking it around the platform, chasing it, and colliding with it again (*Figure 3B*). The movement of the agent ball in the animate condition was manually specified to approximately simulate the movement of an animate, goal-directed agent, while movements of the patient ball were rendered using the Bullet physics simulation engine.

Following the priming clip, participants in both animacy conditions viewed a test clip containing the same platform and assembled tower of blocks. In the test clip, the inanimate patient ball rolls into view headed straight for the tower. Shortly after, the agent ball rolls into the frame along the same trajectory and collides with the patient ball, which proceeds to crash into the tower of cubes, bringing it crashing down. Importantly, both balls in the test clip were rendered according to a physics simulation, and the test clips were identical across animate/inanimate conditions. (*Figure 3C*).

Participants were asked to endorse two statements "*The* [agent color / patient color] *ball caused the tower to fall.*" and "*If the* [agent color / patient color] *ball had not been there, the tower would have remained standing.*" using a slider scale ranging from 0 ("totally disagree") to 100 ("totally agree"). Question order was counterbalanced.

Results

We first investigated whether the animacy manipulation impacted judgments of counterfactual dependence. Participants in the Inanimate Agent group provided low counterfactual dependence ratings for both agent (*M* = 30.4, *SD* = 31.4) and patient (*M* = 28.7, *SD* = 30.8) balls. In contrast, participants gave the highest counterfactual dependence ratings for the agent ball in the Animate Agent group (*Figure 4B*). Statistically, agent animacy significantly

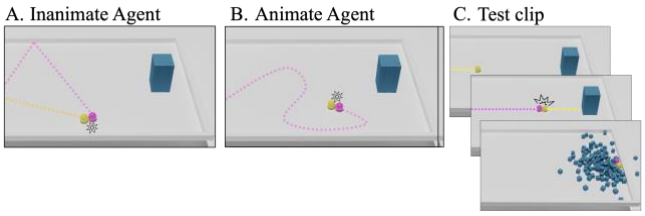

Figure 3: **Experiment 2**. A. Inanimate prime clip. B. Animate prime clip. C. Test clip viewed in both conditions.

affected counterfactual dependence judgments such that the outcome was judged more counterfactually dependent on animate agents (*M* = 51.0, *SD* = 33.19) than inanimate agents (*M* = 30.4, *SD* = 31.4), despite both causes following identical trajectories in the test clip, $F(1, 208) = 21.39$, $p < .0001$.

Given that we succeeded in finding a case where counterfactual judgments come apart for animate and inanimate agents, we next asked whether we also observed differences in causal judgments. Using one-way analysis of variance to test for the total effect of animacy on agent causal ratings, we found that perceived animate agents were indeed rated significantly more causal than inanimate agents for the same outcome event, $F(1, 208) = 10.54$, $p = .00136$ (*Figure 4A*, light plots). Interestingly, causal ratings were highest for the patient ball in the Inanimate Agent condition, suggesting that something other than counterfactual dependence may be driving causal judgments when both balls are inanimate and the outcome is overdetermined.

Finally, we more directly considered the relationship between causal and counterfactual judgments. Controlling for animacy ratings, there was a strong effect of counterfactual dependence judgments on causal rating $F(1, 208) = 72.55$, $p < .0001$. Using the mediation package in R, we also found a significant indirect effect of animacy on causal ratings of the agent ball ($\beta_{animacy} = 20.64$)*($\beta_{counterfactual} = .542$) = 11.19; bootstrapped 95% *CI* [6.077, 16.58], $p < .001$.

Discussion

This study examined whether perceived animacy affects causal judgments by influencing the availability of difference-making counterfactuals. We reasoned that the perception of animacy elicits counterfactuals in which an agent makes a difference in the outcome (*i.e.*, by having a different intention). We found that, when the agent ball was

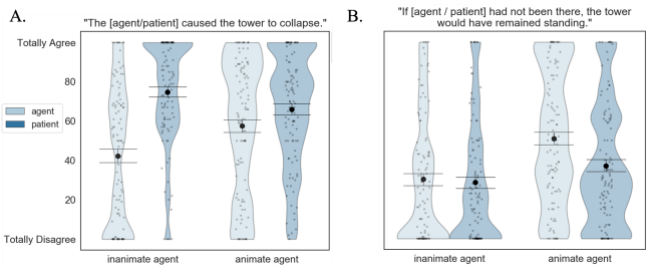

Figure 4: **Experiment 2**. A. Counterfactual dependence ratings to each ball in both agent conditions. B. Causal ratings to each ball in the inanimate and animate agent conditions.

viewed as inanimate, judgments of counterfactual dependence for both causes were low. This result makes sense given that the outcome under consideration was inevitable in this condition and would still obtain in either counterfactual in which one of the objects was removed. When the agent ball was viewed as animate, however, judgments of counterfactual dependence for the agent ball were significantly higher than the inanimate condition, despite participants viewing identical outcomes in the test clip across groups. This could suggest that, in this condition, the outcome may not have been viewed as inevitable. Instead, a relevant counterfactual exists in which an animate agent could have brought about a different result, preserving the tower. A limitation of our design was that we did not assess this possibility more directly, and instead used a counterfactual statement concerning the *absence* of the agent, rather than alternative behaviors. Still, participants may have charitably interpreted the statement along these lines, given the observed patterns of counterfactual judgments.

Experiment 3: Prescriptive norms

Experiment 2 found that animacy *does* impact causal judgments, but only to the extent that it dictates which counterfactual possibilities might be considered. Here, we extend this connection by considering cases of *prescriptive norm* violations (Henne et al., 2019; Knobe, 2009). Using norm violations to study causal cognition is useful because they only apply to animate agents and because prescriptive norm violations make salient counterfactuals in which the violated norm is, instead, followed (Petrocelli et al., 2011). We manipulated both animacy and norm-conformity to examine how they interact in affecting causal judgments of agents in overdetermined cases.

Methods

Participants 587 Participants were recruited in two cohorts from Prolific: 293 adults (M_{age} = 33, *SD* = 12, 148 female, 145 male) were recruited to participate in the causal judgment cohort, and 294 adults (M_{age} = 32, *SD* = 12, 146 female, 148 male) were recruited to the counterfactual judgment cohort.

Stimuli and Procedure Participants in the Inanimate condition viewed a priming clip in which a green and a pink ball rolled into view on a platform that also contained an assembled tower of blocks (as in Experiment 2). The balls collided with each other and bounced off the edges of the platform. Other participants were assigned to one of two animate conditions and viewed a priming clip in which *both* the balls were animated to appear self-guided and goal-directed. The green "builder" ball was shown assembling the tower one cube at a time, while the pink "guard" ball moved interactively as if observing the green ball build the tower. The two animate conditions differed only in the prescriptive norm information provided: In the Immoral condition participants were told that "*It is Pink's job to protect Green's tower*"; in the Irrational condition, they were told that "*Green wants to protect its own tower*".

Participants in all conditions viewed a test clip similar to that used in Experiment 2. It depicted the patient ball rolling into view headed straight for the tower of blocks; shortly after, the agent ball rolls into the frame along the same trajectory and collides with the patient ball, which then crashes into the tower. The test clips were constructed so that participants in the Immoral condition watched the pink "guard" collide with the green "builder", violating the moral prescription to protect the tower. Participants in the Irrational condition watched the green "builder" collide with the pink patient ball, violating the rational prescription to protect its own tower. In the Inanimate cases, both balls were inanimate and the color of the agent vs. patient ball was still counterbalanced but irrelevant. Participants made causal/counterfactual ratings as the test clip looped.

Causal and Counterfactual Judgments Participants in the causal cohort indicated their agreement with the statements, "[Agent/Patient] *caused the tower to fall*". In the counterfactual cohort, participants indicated their agreement with three statements: (1) "*If* [Agent/Patient] *had not been there, the tower would have remained standing*.", (2) "*I expected* [Agent] *to move in a different way than it did in the video*.", and (3) "*If* [Agent] *had moved in a different way, the tower would have remained standing*." All ratings were provided on a slider with values ranging from 0 ("Disagree") to 100 ("Agree").

Results

Causal Judgments Despite having the same underlying physics across conditions, causal attributions to the agent ball significantly differed between conditions, $F(2, 289) = 20.1$, $p < .0001$ (*Figure 5*, light plots). Pairwise comparisons revealed that the agent ball in the Immoral condition was judged as more causal than the Irrational, $t(289) = 5.36$, $p < .000$, and the Inanimate agent, $t(289) = 5.61$, $p < .0001$. Interestingly, causal judgments of the Irrational and Inanimate agents did not significantly differ, $t(289) = 0.250$, $p = .97$. Causal attribution to agents and patients also differed from each other within each condition. In the Inanimate condition, the patient, which makes contact with the tower,

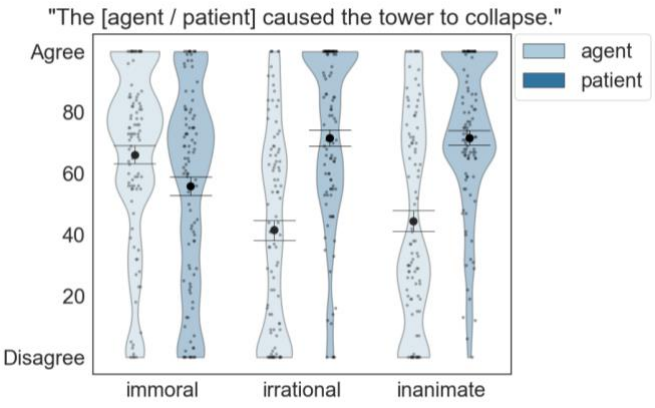

Figure 5: **Experiment 3**. Causal judgments to agent and patient balls across conditions.

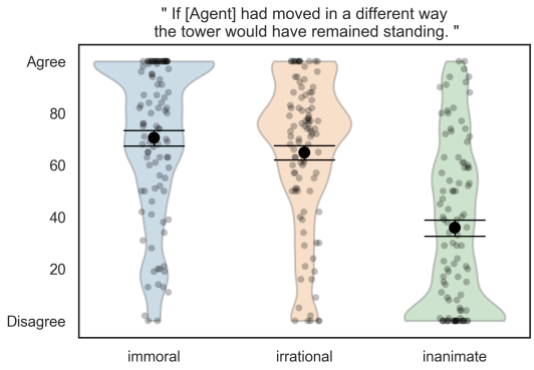

Figure 6: **Experiment 3**. Counterfactual judgments for agent balls across conditions.

was judged more of a cause than the agent, as expected, $t(96) = 6.03$, $p < .0001$, paired (*Figure 5*, right). Though not predicted, this was also true in the Irrational condition, $t(96) = 5.25$, $p < .0001$, paired (*Figure 5*, center). As hypothesized, however, this pattern reverses in the animate Immoral condition, where the agent is judged more causal than the patient (*Figure 5*, left). However, this difference was not significant $t(96) = -1.83$, $p = .06$, paired.

Counterfactual Judgments We found a significant interaction effect of condition × ball (agent / patient) on counterfactual dependence judgments $\chi^2(2) = 54.31$, $p < .0001$. This was driven by the Irrational condition, in which participants judged the tower collapsing as more dependent on the patient ($M = 67.98$, $SD = 35.2$) than the agent ($M = 34.5$, $SD = 34.4$), $t(297) = -7.56$, $p < .0001$. We speculate on why this occurred, and its implications in the discussion section for this experiment. We did not find a significant difference in counterfactual dependency judgments between the agent and patient for the Immoral ($t(297) = 1.552$, $p = .63$) or Inanimate ($t(297) = 1.73$, $p = .52$) conditions.

We also asked participants to judge the extent to which the agent violated their expectations. For this item, Animate agents ($M_{immoral} = 78$, $SD = 26$; $M_{irrational} = 71$, $SD = 28$) were found more surprising than inanimate ones ($M_{inanimate} = 30.19$, $SD = 29.42$), $F(1, 292) = 160.31$, $p < .0001$. There was no difference in the extent to which "behavior" violated participant expectations between immoral and irrational agents $t(291) = 1.75$, $p = .190$.

Finally, to test if participants believed the outcome counterfactually depended on the agent's surprising movement, we asked how much they agreed that the tower's outcome would be different if the agent had moved differently. Results for this item mirrored those of the preceding surprise judgments. Participants agreed more that the outcome would be different if Animate agents ($M_{immoral} = 70.54$, $SD = 29.27$; $M_{irrational} = 63.07$, $SD = 28.75$) moved differently than if the inanimate agent had ($M_{inanimate} = 35.37$, $SD = 31.93$), $F(1, 292) = 73.65$, $p < .0001$. There was no difference in this judgment between immoral and irrational agents $t(291) = 1.51$, $p = .288$ (*Figure 6*).

Discussion

In this experiment, we manipulated the perceived animacy as well the normativity of a causal agent and found that animated "immoral agents" were judged as more causal of an outcome than inanimate objects, even when the actual physical events that occurred were held perfectly fixed. Surprisingly, causal attribution ratings for "irrational agents" appeared similar to those for inanimate objects. One likely explanation is that participants did not actually perceive the animated "builder" agent as acting with the *intention* of making the tower collapse, but rather as failing to prevent the other ball from colliding with it. This explanation is strongly supported by participants' unexpectedly high agreement rating that the outcome was more counterfactually dependent on the patient ball than the "irrational" agent. Thus one possibility is that our stimuli unintentionally demonstrated an interesting dissociation between deliberate and unintentional prescriptive norm violations (*cf.* (Kirfel & Phillips, 2023)).

Overall, the general pattern of causal judgments fits well with the pattern observed for counterfactual dependence judgments in a separate cohort. Taken together, these results provide support for a unified system of causal reasoning, where causal judgments of both animate and inanimate agents are shaped by available counterfactual alternatives.

General Discussion

Experiment 1 asked whether merely being animate was sufficient to alter participants' causal judgments. We found that, when holding both realized and counterfactual outcomes fixed, goal-directed animacy had no effect on causal judgments. Instead, we replicated prior work demonstrating that causal attribution to agents, and objects alike, is a function of counterfactual difference-making (Danks, 2017). Experiment 2 investigated cases of overdetermination in which animacy allows for a dissociation between the availability of difference-making counterfactuals. Here, we observed differences in causal judgments that were mediated by judgments of counterfactual dependence. Experiment 3 tested the effect of norm violations in a similar causal structure and found higher causal attributions to a perceived "immoral agent" who violated a prescriptive norm but did not find higher causal attributions to an "irrational" agent who was likely perceived as acting unintentionally.

In this set of experiments, we found that animacy, itself, does not make agents more causal of an outcome than objects. Instead, causal judgments about agents and objects differ as a function of the counterfactuals they respectively afford. Finally, the counterfactuals used to make causal attributions to agents and objects are distinguished by normative expectations for how they should behave. A unified account of causal reasoning may come from future work exploring normative expectations as a domain-general bridge between intuitive physics and folk psychology.

References

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, *63*(3), 368–378.

Bishop, J. M. (2020). Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It. *Frontiers in Psychology*, *11*, 513474.

Danks, D. (2017). Singular causation. *The Oxford Handbook of Causal Reasoning*, 201–215.

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, *7*(1), 1–9.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936–975.

Henne, P., Niemi, L., Pinillos, Á., De Brigard, F., & Knobe, J. (2019). A counterfactual explanation for the action effect in causal judgment. *Cognition*, *190*, 157–164.

Henne, P., O'Neill, K., Bello, P., Khemlani, S., & De Brigard, F. (2021). Norms affect prospective causal judgments. *Cognitive Science*, *45*(1), e12931.

Hilton, D. J., McClure, J., & Moir, B. (2016). Acting knowingly: effects of the agent's awareness of an opportunity on causal attributions. *Thinking & Reasoning*, *22*(4), 461–494.

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

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

Kirfel, L., & Lagnado, D. A. (2018). Statistical norm effects in causal cognition. *CogSci*. https://cogsci.mindmodeling.org/2018/papers/0132/0132.pdf

Kirfel, L., & Phillips, J. (2023). The pervasive impact of ignorance. *Cognition*, *231*, 105316.

Knobe, & Fraser. (2008). Causal judgment and moral judgment: Two experiments. *Moral Psychology*. https://files.osf.io/v1/resources/5eanz/providers/osfstorage/591f41446c613b024dd1e033?action=download&direct&version=1

Knobe, J. (2009). Folk judgments of causation. *Studies in History and Philosophy of Science. Part B. Studies in History and Philosophy of Modern Physics*, *40*(2), 238–242.

Kominsky, J. F., & Phillips, J. (2019). Immoral Professors and Malfunctioning Tools: Counterfactual Relevance Accounts Explain the Effect of Norm Violations on Causal Selection. *Cognitive Science*, *43*(11), e12792.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2022). Emmeans: Estimated marginal means, aka least-squares means. *R Package Version*.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A Theory of Blame. *Psychological Inquiry*, *25*(2), 147–186.

Michotte, A. (1946). *La Perception de la Causalité*. Inst. Sup. De Philosophie.

Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of Personality and Social Psychology*, *100*(1), 30–46.

Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition*, *156*, 164–176.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind Games: Game Engines as an Architecture for Intuitive Physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, USA.

Wu, S., Sridhar, S., & Gerstenberg, T. (2022). That was close! A counterfactual simulation model of causal judgments about decisions. *Cognitive Science Proceedings*.