Spring 6-10-2023

Epistemic Mentalizing and Causal Cognition Across Agents and Objects

Bryan S. Gonzalez
Dartmouth College, bryan.s.gonzalez.gr@dartmouth.edu

EPISTEMIC MENTALIZING AND CAUSAL COGNITION ACROSS AGENTS AND OBJECTS

A Thesis

Submitted to the Faculty
in partial fulfillment of the requirements for the
degree of

Doctor of Philosophy

In

Cognitive Neuroscience

by

Bryan S. Gonzalez

Guarini School of Graduate and Advanced Studies
Dartmouth College
Hanover, New Hampshire
April 2023

Examining Committee:

_____
(chair) Jonathan S. Phillips, Ph.D.

_____
Meghan L. Meyer, Ph.D.

_____
Mark A. Thornton, Ph.D.

_____
Tobias Gerstenberg, Ph.D.

_____
F. Jon Kull, Ph.D
Dean of the Guarini School of Graduate and Advanced Studies

Abstract

This dissertation examines mentalizing abilities, causal reasoning, and the interactions thereof. Minds are so much more than false beliefs, yet much of the existing research on mentalizing has placed a disproportionately large emphasis on this one aspect of mental life. The first aim of this dissertation is to examine whether representing others' knowledge states relies on more fundamentally basic cognitive processes than representations of their mere beliefs. Using a mixture of behavioral and brain measures across five experiments, I find evidence that we can represent others knowledge quicker and using less neural resources than when representing others' beliefs. To be considered a representation of knowledge rather than belief, both mentalizer and mentalizee must accept the propositional content being represented as factive (Kiparsky & Kiparsky, 2014; Williamson, 2002). As such, my results suggest that representing the mental states of others may be cognitively easier when the content of which does not need to be decoupled from one's own existing view of reality.

Our perception of other minds can influence how we attribute causality for certain events. The second aim of this dissertation is to explore how perceptions of agency and prescriptive social norms influence our intuitions of how agents and objects cause events in the world. Using physics simulations and 3D anthropomorphic stimuli, the results of three experiments show that agency, itself, does not make agents more causal to an outcome than objects. Instead, causal judgments about agents and objects differ as a function of the counterfactuals they respectively afford. Furthermore, I show that what distinguishes the counterfactuals we use to make causal attributions to agents and objects are the prescriptions we hold for how they should behave.

Why do we say a fire occurred because of a lightning strike, rather than the necessary presence of oxygen? The answer involves our normative expectations of the probability of lightning strikes and the relative ubiquity of oxygen (Icard et al., 2017). The third aim of this dissertation explores the gradation of causal judgements across multiple contributing events that each vary in their statistical probability. I

contribute to ongoing theoretical debates about how humans select causes in experimental philosophy and cognitive science by introducing a publicly available dataset containing 47,970 causal attribution ratings collected from 1,599 adult participants and structured around four novel configurations of causal relationships. By quantitatively manipulating the influence of normality, I systematically explore the continuous space of a causal event's probability from unlikely to certain. It is my hope that this benchmark dataset may serve as a growing testbed for diverging theoretical models proposing to characterize how humans answer the question: Why?

Acknowledgements

This dissertation would not be possible without the support of several important individuals. First, I would like to acknowledge the advice and support I have been lucky to receive from my faculty committee. Mark, your multilayered view of the social mind has directly influenced my thinking and fascination with mentalizing. Tobi, your models of causal cognition have been pivotal to so many ideas in this document and to the questions that keep me up at night. Thank you for inspiring me. Meghan, your compassion has been a revelation that success as a scientist and genuine human empathy are not mutually exclusive. When the hurdles of graduate school are over, and my dmPFC earns a moment to rest, I'll remember your kind encouragement. Thank you for seeing me, not as another stressed graduate student, but as a human.

To my advisor, Jonathan Phillips, thank you for accepting a disillusioned, cynical, 5th-year graduate student to your team. In moments when I doubted whether the path to this point was worth the pain, your real and concerted efforts to foster a welcoming environment and hold space for every voice in the room have made all the difference in my choice to push forward. I grew more confident with each passing week in your lab that I had finally found a place where I socially and intellectually belonged. You have expanded my mind in ways that made me genuinely excited to do this work. Your approach to asking deep questions makes working together feel like a pursuit of my genuine passions and curiosity, rather than mere drudgery for the sake of an "impressive" CV or career. This sense of intrinsic motivation led me to graduate school in the first place and is the reason I have been able to finish it. Thank you.

Years ago, a graduate student changed my trajectory by communicating her science in a vocabulary I understood as an outsider. I was "let in" to academic research at that moment, could share in her excitement for it, and gladly slogged through the tedium of an assistant because I was shown the bigger picture of *why* it mattered. Thank you, Dr. Lauren Aguilar. My time at Dartmouth has had purpose because I've been able to hold the door open behind me, letting others "in" to science who, like me, were

at the margins. I'm grateful to the brilliant undergraduates who have helped me complete this work in different ways: Karina Lopez '19, Brandon Dormes '23, Orlando Valladares '26. *¡Pa'lante, siempre pa'lante!*

To fellow graduate students, Manish, Sushmita, thank you for every coffee break, every walk around Occom, and every hallway chat. These moments sustained me. To my cohort for commiserating through the chaos and celebrating each others' successes. To Vassiki Chauhan and Sasha Brietzke, thank you for being my fighters and friends.

I'm grateful for the support of my tribe back home in New York. To Leo Benton, for teaching me to pursue a life of depth over frivolous superficiality. To Kieran Baker, for making my goals your own, and showing me unconditional support. To my day ones, Ashley Lekwauwa and Heather Dumorne. Thank you for growing into adulthood with me - unlearning what we thought mattered and learning to survive together. Your presence throughout the last six years has fed my soul in more ways than you know.

To my brother, Jonny, and sister, Kat, thank you for continually showing me that *real* success is in honoring every hurdle and showing grace to others and myself. At times when the ivory tower made me resent our humble roots for not grooming me for the elite, your unconditional love and support kept me tethered to what matters most and helped me to see true strength in what I once viewed as weakness. I love and admire you both.

To Mert Özkan. Beyond my gratitude for your unwavering friendship, providing relief from the isolation of graduate school, I am grateful for the gift of who you are. From your righteous convictions and hilarious sense of humor to your terrible driving and delicious cooking. Our connection as scientists and, more importantly, as imperfect humans, has been my truest source of comfort through every rough patch of the journey. In the desolate, frozen woods, thank you for being my home. I love you.

To my mother, Irma. My successes will always be in honor your American dream. Thank you for showing me what hard work can accomplish. For giving me your determined ambition so that, just like my hero, I could cross any border in my path.

Contents

List of Figures

List of Tables

Chapter 1.

"*Be patient toward all that is unsolved in your heart and try to love the questions themselves,*
like locked rooms and like books that are now written in a very foreign tongue.
Do not now seek the answers, which cannot be given you
because you would not be able to live them.

And the point is, to live everything. Live the questions now.
Perhaps you will then gradually, without noticing it,
live along some distant day into the answer."
- Rainer Maria Rilke

1.1 Overview

Our mental representations of another person involve a unique configuration of causal relationships we can use to explain, influence, and predict their behavior. These cause-effect relationships are part of a network of associations that includes links between another person's internal mental state and externally observable behavior. Despite this division between internal and external worlds, these links allow us to mentally reverse-engineer others' behavior to deduce the hidden mental states from which they arise.

The projects in this dissertation sit at the intersection of theory-of-mind reasoning (chapter 2) and causal reasoning (chapter 4) and explore the interaction between them (chapter 3). More specifically, the second chapter of this dissertation addresses the existing debate in the mentalizing literature about whether the ability to represent what other agents know relies on more fundamentally basic cognitive processes than the ability to represent what other agents believe. The third chapter of this dissertation considers the role of theory-of-mind in the causal reasoning process and specifically asks how inferences about an agent's mental state might alter our causal ascriptions by influencing the counterfactuals that

could be considered for outcomes caused by intentional or unintentional agents as compared to inanimate objects. Finally, the fourth chapter of this dissertation contributes directly to the study of causal reasoning, focusing specifically on the way in which statistical information about events affects judgments of causal selection.

1.2 Knowledge without Belief

The capacity to form theories of other minds has been said to rely on commonsense intuitions of how the environment causes mental states and how those mental states further cause behaviors (Carruthers & Smith, 1996; David Lewis, 1972). We traverse these causal paths when inferring mental cause from behavioral effect or, moving in the opposite direction, using the inferred mental state of another person to predict the behaviors that follow. This ability to perceive the hidden mental causes of overt behavior has endowed our species with the remarkable ability to cooperate and compete (Adolphs, 2009; R. Dunbar, 2003; R. I. M. Dunbar, 1998). However, we are still far from fully understanding how this 'mind reading' ability develops and operates. A stringent test for the capacity to mentalize has been proposed by philosophers for the presence of mentalizing abilities, requiring the prediction of another person's behavior on the basis of that person's false beliefs (Bennett, 1978; Dennett, 1978). Proponents of this criteria argue that predicting behavior caused by true beliefs would be insufficient evidence of a capacity for mentalizing since it would be impossible to discern whether the prediction is in accordance with reality, or in accordance with another person's inferred beliefs about reality. Thus, the study of mentalizing became the study of predicting behavior from false beliefs (Baron-Cohen, 1997; Call & Tomasello, 2008; C. D. Frith & Frith, 2012; Wimmer & Perner, 1983). Although valuable insight has been gleaned from classic false-belief tasks, a paradigm shift is needed in which mentalizing is studied in ways that better reflect how we typically use these abilities in daily life. That is, not in the representation of others' uncertain beliefs, but in the representation of others as sources of genuine knowledge.

1.2.1 The difference between knowledge and belief

Multiple features of knowledge distinguish it from mere beliefs. The representation of another person's knowledge state, for instance, is constrained by what one takes to be true (Kiparsky & Kiparsky, 2014; Williamson, 2002) (e.g. the mentalizer cannot represent another person as knowing that *Santa Clause actually delivers gifts*; unless the mentalizer accepts the proposition, *Santa Clause actually delivers gifts*, as true.). On the other hand, representations of others' *beliefs* are unconstrained by the facts and free to contain anything imaginable (e.g. little Suzy believes a man named Santa Clause uses flying reindeer to deliver gifts). This condition of factivity in knowledge, does not, however, preclude the representation of others as possessing knowledge that a mentalizer, themselves, lacks (Karttunen, 1977; Phillips & George, 2018). We can represent John as knowing the directions to the wedding venue without knowing, ourselves, how to get there. This *egocentric ignorance* highlights the way knowledge and beliefs also differ in their propositional logic. For instance, when an agent does not know some proposition, *p*, they are agnostic as to whether *p* is true or false. On the other hand, when an agent does not *believe* some proposition, *p*, they take *p* as decidedly false (e.g. Billy does not *know* if Santa is real vs. Billy does not *believe* that Santa is real) (Laurence R. Horn, 1989; R. T. Lakoff, 1968). Importantly, many instances of even *true* beliefs fail to meet the criteria for knowledge. Examples of this come from "Gettier cases", in which a person forms a belief on the basis of evidence that falls short of certainty, but happens to be true by coincidence (Chisholm, 1966; Gettier, 1963; Machery et al., 2017; Starmans & Friedman, 2012). For instance, Billy sees what looks exactly like a sheep in the hills. From this, he is justified in his belief of the proposition, *there is a sheep in the hills.* Now consider that what Billy sees is actually a dog disguised as a sheep. Also consider that there *is*, in fact, a genuine sheep in the hills that Billy cannot see. In this case, Billy's belief that *there is a sheep in the hills* is true, but one would not intuitively represent Billy as *knowing* this fact (Chisholm, 1966).

1.2.2 Comparative and developmental evidence

A more complete understanding of human cognition can be achieved by first defining which of its fundamentally basic capacities are recruited in service of more sophisticated abilities. Comparative studies of non-human primates, as well as developmental research using children, provide useful ways of disentangling the core aspects of mentalizing from its emergent complexity in more sophisticated forms.

In studying the development of human mentalizing abilities across evolutionary history, we can track which cognitive abilities are preserved backward in time along human phylogeny. It stands to reason that more basic cognitive processes will be present in species more distal to humans along the evolutionary continuum. Research on the presence of false-belief representations in our close phylogenic relatives, the great apes, remains ambiguous. In a looking paradigm that manipulated knowledge and false beliefs, Kano and colleagues found that chimpanzees could predict where human actors behind and occlusion will search for an object on the basis of their false belief about its location (Kano et al., 2019; Martin, 2019). Although compelling, other groups have found conflicting results. In a decision-making context, Kaminski et. al. report that chimpanzees failed to capitalize on a dominant competitor chimp's false belief about the location of a food reward (Kaminski et al., 2008). Taken together, the controversy of whether great apes can represent a conspecific's beliefs remains theoretically unsolved.

Despite these mixed results, great apes demonstrate unequivocal evidence for an ability to mentally represent what another agent knows. In an earlier version of the competitive social decision-making task used by Kaminski et al., Hare and colleagues found that subordinate chimpanzee behavior was, in fact, sensitive to the information known by a dominant competitor. In this paradigm, if a subordinate chimp could see that its dominant competitor knew the location of a hidden food item, it showed no signs of reaching for it (Hare et al., 2000). Remarkably, if the knowledgeable dominant chimp was replaced with another dominant chimp that was instead ignorant to the food's location, the subordinate subject preferentially approached and more often retrieved the hidden food reward (Hare et al., 2001). Overall, prior work investigating mentalizing abilities in great apes has demonstrated their genuine ability to represent what other agents know but has reached far less certain conclusions about their ability to attribute beliefs (Bräuer et al., 2007; Karg et al., 2015; Krachun et al., 2009).

Traveling further down the evolutionary tree, research on more phylogenetically distant monkey species paints a clearer picture of the limits of their mind-reading abilities. This work has shown monkeys to possess the distinct capacity to represent what others know, but not what others categorically believe (Drayton & Santos, 2018; Flombaum & Santos, 2005; Martin & Santos, 2016).

We have learned a great deal about the sophisticated mentalizing abilities of adults from research investigating its precursors in developing children. Classic work on the representation of false beliefs in children using the famous Sally-Anne task, which meets the aforementioned criteria of predicting an agent's behavior based on their false beliefs, has established unequivocal benchmarks for the age in which more complex mentalizing abilities develop from simpler ones. This work demonstrates that children do not possess the ability to represent another agent's false belief before the age of four (Baron-Cohen et al., 1985). Interestingly, there is evidence to suggest that, between the ages of four to six, children further fail at representing justified true beliefs like the ones described in the Gettier case above (Oktay-Gür & Rakoczy, 2017).

In comparison to the late emergence of belief attributions, work on knowledge attribution in younger populations indicates that these abilities exist much earlier in human development. Infants as young as six months old have demonstrated a sensitivity to whether or not others agents have perceptual access to preferred objects (Luo & Johnson, 2009). Even in the development of language, children learn the meaning of verbs used to ascribe knowledge ("knows", "understands") earlier than those used to attribute beliefs ("thinks", "assumes") (Moore et al., 1989).

1.2.3 Evidence from clinical populations.

Studies of people with clinical diagnoses can also reveal which processes serve as fundamentally core aspects of cognition as we would expect these processes to be preserved in these populations while more sophisticated abilities are disrupted. People living with Autism Spectrum Disorder (ASD) have been of particular interest in research on mentalizing because they show a relatively consistent pattern of symptomology in specific aspects of social cognition. Foundational work has found that a hallmark of

ASD is a delay in the ability to pass false-belief tasks (Baron-Cohen, 1997; Baron-Cohen et al., 1985; U. Frith, 2001; Moran et al., 2011; Senju et al., 2009). Despite these challenges, other findings reveal an ability comparable to neurotypicals in attributing factive mental states to others given their perceptual access (Hobson, 1984), and desires (Baron - Cohen, 1989).

Evidence for the ability of individuals with autism to attribute factive mental states to others is not limited to contexts involving an agent's mere line of vision, although these instances do provide sufficient justification for perceptual access to constitute knowledge (Lyons, 2017). Linguistic approaches have also shown that autistic individuals can intuit a speaker's knowledge from certain kinds of implicatures. Consider the statement "some of the students passed the exam". According to popular accounts of communication, the understanding from this statement that not *all* students passed requires the assumption that the speaker would have uttered a stronger/more informative alternative if they knew it to be true (Gazdar, 1979; Grice, 1969; Laurence Robert Horn, 1972). As such, these implicatures require epistemic reasoning about others (Sauerland, 2004). Hochstein and colleagues found that adolescents with ASD did not differ from typically developing participants in comprehension of these kinds of conversational implicatures (2018).

1.2.4 The current investigation

The purpose of experiments reported in chapter 2 of this dissertation is to determine whether the mental representation of another agent's knowledge requires the additional representation of their beliefs, incurring the cognitive costs thereof. A simple way of determining whether people can make evaluations of knowledge in the absence of evaluations of belief is to investigate the speed with which these evaluations are made. I begin with a simple task that investigated the response times for evaluations of knowledge and belief ascriptions. Participants read about various people in a variety of different scenarios and then made truth value judgments of knowledge and belief ascriptions to those agents. If participants' evaluations of knowledge require or involve evaluations of belief, we would expect evaluations of knowledge ascriptions to be slower than evaluations of similar belief ascriptions.

One potential concern with this initial study is that it involves cases in which some of the ascriptions of belief may be pragmatically odd or infelicitous. Specifically, it may sound unnatural to describe an agent as believing some proposition in cases where the agent meets all of the criteria for knowledge. Our design does just this for methodological consistency. In a second experiment, I test for the impact of pragmatic infelicity while also asking whether differences in response times for evaluations of others' knowledge and belief generalize to participants with ASD who vary in understanding linguistic pragmatics that require reasoning about others' beliefs. I measured the extent to which the mental state ascriptions participants evaluated were perceived as pragmatically odd. If pragmatic differences in language are responsible for a comparative delay in evaluations of others' knowledge or belief, we should expect these differences to be reflected in the response time patterns of the ASD group more than the neurotypical control group. Alternatively, if a response time discrepancy instead resulted from knowledge assessments being made in the absence of calculations of others' beliefs, we expect participants with autism to demonstrate the same, or more extreme, differences in response times between knowledge and belief ascriptions.

I next ask whether a response time difference between evaluations of knowledge and belief exists in languages other than English by conducting a highly similar experiment in French. French provides a particularly strong test because, unlike English, the French term used for belief ascriptions is used roughly 1.49 times more frequently than the French term used for knowledge ascriptions. Thus, if differences in response times merely reflect lexical frequency, we should expect evaluations of beliefs to be faster than evaluations of knowledge in French.

In a fourth experiment, I consider the broader class of factive and non-factive mental state attitudes. Factive mental states include those of 'observing', 'understanding', and 'recognizing'. On the other hand, non-factive mental states include those of 'assuming', 'imagining', and 'predicting'. I conduct a similar experiment, but replace the words 'know' and 'think' with counterparts from their respective factive and non-factive class of verbs. A plausible reason to predict this general difference is that factive mental state representations may be simpler because the content represented is necessarily consistent with

one's own understanding, and thus does not have to be represented separately (Phillips & Norby, 2019). If this prediction proves true, it could point toward a more general explanation of why people are faster to evaluate knowledge ascriptions than belief ascriptions.

In the final experiment of chapter 2, I ask a similar question to the general one tested across the previous four experiments but use a very different methodology. Specifically, I use functional magnetic resonance imaging (fMRI) to examine the neural responses exhibited during the formation of knowledge and belief representations and ask whether these neural patterns provide evidence about the relationship between knowledge and belief that converged with my prior findings.

1.3 Causal Cognition for Agents & Objects

Research on causal cognition for physical and social events has, thus far, been pursued as distinct endeavors. In what follows, I review previous work on the mechanisms of causal reasoning for objects and agents. I then describe how this dissertation will demonstrate a more cohesive view of causal cognition across animate agents and inanimate objects.

1.3.1 Philosophical perspectives

Much of the philosophical work on causality has attempted to define the conditions that qualify things or events as causes. The numerous proposals to describe the properties that constitute causality have fallen into two broad classes. Process theories of causality require that causes be linked to their effects by virtue of a spatiotemporally continuous process in which some quantity, such as physical force, is transferred from cause to effect (Aronson, 1971; P. Dowe, 2000; Fair, 1979; Machamer et al., 2000; Salmon, 1984, 1994; Waskan, 2011). For example, a person who throws a rock is said to cause a window to shatter by virtue of the spatiotemporally continuous process in which the force generated by the thrower is transferred to the resting glass window. On the other hand, dependence theories of causation claim that a candidate event is causal if an outcome event's occurrence is dependent on the candidate in

some way. This dependence can be demonstrated by appealing to counterfactuals such that some candidate, *c*, is the cause of an outcome, *e*, if *e* would not have occurred in the counterfactual where *c* is absent (David Lewis, 1974; Mackie, 1980). Returning to the broken window example, the person who threw the rock qualifies as a cause of the window breaking according to dependence theories since the same outcome would not occur in a counterfactual that omits the event of the person throwing a rock.

While both process and dependence frameworks make similar causal attributions in this simple case involving one candidate cause and one outcome, more complicated cases can call each theory into question. In cases of overdetermination, two events occur that are both sufficient to independently cause an outcome. Consider that person A and person B both throw rocks that hit the same window at precisely the same time. In this case, both candidates demonstrate causality according to process accounts (they both transfer a force to the window), but neither candidate demonstrate counterfactual dependence (i.e. in the counterfactual where person A is omitted, the window still breaks due to person B). An alternative causal structure in which process and dependence accounts make different conclusions is known as double prevention. To illustrate, consider a scenario in which person A fires a gun to assassinate a target. A bodyguard rushes to intercept the bullet, but a bystander accidentally trips him, preventing him from preventing the target's death. Process theories would argue that the gunman caused the target's death since their action transferred a quantity of force to the target in a spatiotemporally continuous way. Dependence theories, however, would claim that the bystander was also a cause of the target's death since the target would have lived in a counterfactual in which the bystander had not tripped the bodyguard. Scenarios involving overdetermination demonstrate causality by satisfying process, but not dependence, theories. While scenarios involving double prevention demonstrate causality from dependence without meeting the criteria of process theories (Hall, 2004). The discrepancy in these hypothetical cases highlights the complexity of causal reasoning that exists in the real world in fields such as medicine and law.

1.3.2 Causal cognition for objects

Humans possess a natural understanding of the physical world and the fundamental principles that govern the behavior of physical systems. This understanding is acquired through everyday experiences and does not require explicit instruction or formal training in physics. (Kubricht et al., 2017; McCloskey, 1983; Ullman et al., 2018). Our intuitive sense of physics allows us to make predictions about how objects will behave based on their properties and the forces acting on them. For example, we can intuitively understand that a ball will roll downhill and that heavier objects fall faster than lighter ones.

Our sense of physics plays a crucial role in our understanding of causality. Causal attribution for physical events has often been studied using stimuli depicting collisions between rigid bodies. In the classic work of Albert Michotte, a moving *agent* appears to collide with a static *patient*, exerting a force onto and subsequently "launching" the patient into motion. In these simple contexts, people overwhelmingly agree that the agent causes the patient's motion (Michotte, 1946).

In recent work, Gerstenberg and colleagues asked participants to make causal and counterfactual judgments of moving shapes presented in scenes depicting various configurations of agent causes and patient outcomes. Their proposed Counterfactual Simulation Model (CSM) provides a computational framework for causal judgments that brings together aspects of both dependence and process theories of causation. The CSM begins by determining which candidates in a scene make a difference in the outcome. It does this using a programmable physics engine injected with small amounts of noise to mimic the way humans rely on an intuitive sense of physics to simulate possibilities (Ullman et al., 2017). By removing a candidate object under consideration from a scene and playing the simulation engine forward, the model can observe if the outcome counterfactually depended on the presence of the candidate object (Hiddleston, 2005; Pearl, 2000; Woodward, 2003). After determining which objects make a categorical difference in the outcome, The CSM moves on to measures for determining the causal strength of each candidate cause. Appealing to dependence theories, "whether-causation" captures the extent to which a candidate made a difference in the outcome. For the CSM and humans possessing imperfect physics simulations, whether-causation is influenced by the ambiguity of counterfactual outcomes. Additionally,

"how-causation" measures the effect of a perturbance to a candidate object on *how* the outcome obtains, revealing whether a transfer of force exists between the object and the outcome. Along with additional measures of sufficiency and robustness, the CSM provides estimates of causal strength that closely approximate the causal and counterfactual judgments made by participants (Gerstenberg et al., 2021).

1.3.3 Causal cognition for agents

Intentions can serve as one way to separate agents from objects. Consistent results have demonstrated that agents who bring about outcomes deliberately are judged as more causal than those who cause the same outcome unintentionally (Fincham & Jaspers, 1980; Lagnado & Channon, 2008; Malle et al., 2014; McClure et al., 2007). However, the methodological considerations of isolating judgments of causal from moral responsibility can be a challenge when interpreting these findings (Kominsky & Phillips, 2019). Existing evidence suggests that when asked to evaluate the morality of actions, people care about an agent's mental state, whereas considerations of blame are more sensitive to an agent's causal role in an outcome (Cushman et al., 2008; Langenhoff et al., 2021; Malle, 2021). Using anthropomorphized agents who caused harmful outcomes to others, Sosa et al. found that the physical causal contributions of an agent computed using the Counterfactual Simulation Model did not significantly improve predictions of moral judgments above those made by a simpler model using only measures of the agent's inferred desire to cause harm (2021).

Unlike objects, the events caused by agents possess a property of *equifinality*, wherein variation in the means can still lead to the same outcome (e.g. there's more than one way to skin a cat). This property may suggest that the counterfactuals we consider when attributing causality to goal-directed agents are not of variations in their behaviors (as in "how-causation" for objects above) but instead over counterfactual intentions or goals an agent could pursue.

1.3.4 The current investigation

The purpose of the experiments reported in chapter 3 of this dissertation is to examine the common underpinnings of causal cognition across animate agents and inanimate objects. Previous work studying causal attribution in the context of inanimate objects has benefitted from the computational tractability of physical possibilities. The laws of physics according to Newtonian mechanics serve to constrain counterfactuals in ways that allow programmable tests of dependence and causal strength. The success of accounts for causal objects like the Counterfactual Simulation Model, however, cannot generalize to the mental causes of human behavior. This is because a mapping from mental states to behavior is rarely one-to-one. Many distinct mental states can result in the same behavioral effects (e.g. tears of joy vs tears of sadness). Additionally, equifinality tells us that many distinct behaviors can result from the same mental state. To narrow the gap between the conceptualization of causal agents and objects, my approach is to study human judgments in contexts where I minimally vary the agentive status of a potential cause while controlling the fine-grain kinematics of other physical dynamics contributing to an outcome.

First, I explore whether causal attributions are impacted by the animacy of a candidate cause. I use realistic video stimuli of collisions in a billiards context consisting of an agent and patient ball in the classic sense such that the agent always collides with the patient ball, launching the patient *into*, or diverting the patient *from* the corner pocket of a billiards table. However, I manipulate the animacy of the agent ball as appearing either animate and goal-directed, or inanimate and moving in accordance with classic laws of physics. Participants make causal judgments with respect to the outcome of the inanimate patient that either lands in or misses a corner pocket of the table. I further manipulate patient counterfactual outcomes. Thus, I use a 2 (animate vs inanimate agent ball) x 2 (patient outcome: in vs out) x 2 (patient counterfactual: in vs out) crossed design. Crucially, fine-grain physical parameters of the collisions, patient outcomes, and patient counterfactuals are held fixed across animacy conditions, allowing us to observe any effect of animacy, in isolation, on causal judgments.

Since agency instills equifinality in the outcomes caused by goal-directed behavior, manipulations of animacy should promote consideration of counterfactuals in which agents possess

different intentions rather than counterfactuals in which they execute different behaviors. Experiment 2 of this chapter explores this mediating influence of counterfactuals on causal judgments of agents and objects. I explore causal judgments in a new setting, in which two balls are each individually sufficient to bring about the outcome. This time, I manipulate only the *perception* of animacy between subjects through a priming video and present all participants with the same outcome, rendered from a physics simulation in a subsequent test video. I hypothesized that subjects who viewed the collision as one between two objects would view the outcome as inevitable, thereby lacking counterfactuals for either ball that lead to a different outcome. However, for subjects who viewed the collision as one between an intentional agent and an inanimate object, a relevant counterfactual *would* exist in which an agent with a different intention could make a difference in the outcome. I test for differences in casual attribution to seemingly goal-directed agents and inanimate objects for the same outcome and ask if any differences result from the influence of counterfactuals.

One variable known to influence judgments of both animate and inanimate causes is normativity (Halpern & Hitchcock, 2013; Hitchcock & Knobe, 2009; Icard et al., 2017; Kominsky & Phillips, 2019; Morris et al., 2019). Consistent evidence has demonstrated that people have a tendency to attribute increased causality to agents that violate social or moral norms as compared to agents who do not (Henne et al., 2019; Kirfel & Lagnado, 2018; Knobe, 2009; Kominsky et al., 2015). In the final experiment of chapter 3, I manipulate both animacy and normativity between participants. I asked whether or not moral norms are special cases for causal cognition by presenting subjects with a moral norm violation as well as a rational norm violation in which an animate agent behaves inconsistently with their expressed interests.

Taken together, this work attempts to bridge the mechanisms of causal cognition across animate agents and inanimate objects. In both of these domains, causal judgments are strongly influenced by the counterfactuals we consider. In turn, the alternative possibilities that come to mind are dictated by the normative expectations we hold from our intuitive sense of physics or our commonsense folk psychology.

1.4 A Benchmark Dataset of Token Causal Selection

Most events in real life occur through the confluence or consequence of many distinct causal conditions. Consider the extensive chain of cognitive and neural events required to coordinate muscle contractions for even the simplest of behaviors. Yet we do not say that a person smiles because efferents from the ponto-medulliary junction carry neural signals through the seventh cranial nerve…etc. Every event in this chain constitutes a cause according to the philosophical theories described above ("the problem of isomorphism" (Halpern & Hitchcock, 2013, p. 415)). However, we might simply explain that a person smiles from the "actual cause" of satisfaction. Causal selection is the process by which humans decide on which of multiple candidate events constitutes the definitive actual cause of an outcome. While we may select satisfaction to be the cause of a smile, we would not deny that the numerous elements within the neural pathways involved are also causal to smiling. This is because causal responsibility for outcomes can be distributed across events in a graded fashion.

When considering what events might be causes, the various links within and across causal conditions to their effects define the structure of a causal system. In conjunctive causal structures, the occurrence of two or more distinct events, jointly, causes an outcome such that each event is necessary, but no event is, alone, sufficient for the outcome to obtain. In disjunctive causal structures, the occurrence of at least one of multiple events can cause an outcome such that each event is sufficient but not individually necessary for the outcome to obtain.

In chapter 4 of this dissertation, I explore how humans rank and select causal events. More specifically, I focus on judgments of *specific* token events (e.g. '*this* smile is caused by *that* joke') rather than more general causal relationships between properties or kinds (e.g. 'jokes cause smiles'). A benchmark dataset of causal judgments is introduced that can serve to arbitrate competing theories of how we select causes.

1.4.1 The pervasive influence of norms

Our normative expectations play an important role in how we attribute responsibility across multiple causal events (Hall, 2007; Halpern & Hitchcock, 2013, p. 415; Hitchcock & Knobe, 2009; Icard et al., 2017; Knobe & Fraser, 2008; Kominsky & Phillips, 2019; Kominsky et al., 2015; Menzies, 2004). In conjunctive causal systems, people consistently attribute more responsibility for an outcome to causal events that are considered rare. This effect is referred to as *abnormal inflation*: the responsibility attributed to one cause in a conjunctive causal system increases as a function of its abnormality relative to other necessary causes (Hart & Honoré, 1985; Hilton & Slugoski, 1986; Icard et al., 2017; Kahneman & Miller, 1986; Kahneman et al., 1982). In contrast, people tend to attribute less responsibility to causal events that seen are relatively more probable than the other necessary antecedents. This pattern is known as *supersession*, wherein the responsibility attributed to one cause in a conjunctive system decreases as a function of its normality relative to other necessary causes (Kominsky et al., 2015).

A different pattern in causal selection emerges when considering disjunctive structures. That is, people consistently attribute less responsibility to causal events that are considered rare. This effect is referred to as *abnormal deflation*: the responsibility attributed to one cause in a disjunctive causal system decreases as a function of its abnormality relative to other sufficient causes (Icard et al., 2017). Additionally, past work on causal selection in disjunctive systems has revealed an *absence or reversal of supersession*, wherein the responsibility attributed to one cause in a disjunctive causal system is unchanged or increases as a function of its normality relative to other causes (Kominsky et al., 2015).

1.4.2 Innovating research on causal cognition

In what follows, I outline the ways in which the benchmark augments existing research on causal selection and provide a brief overview of the methodology used to collect this rich source of data.

As described above, the precise structure of a causal system has major implications for the things we select as actual causes. Whether using vignettes about people and their intentions or simulated interactions of objects under the constraints of physics, prior research has typically operationalized

conjunctive and disjunctive causal structures as existing between just two distinct events leading to a single outcome. These designs have been extremely fruitful to our understanding of causal cognition. More useful insight may be gained, however, by augmenting these structures with more nodes so that they better reflect the causal relationships we encounter in the real world. While it is unlikely that we entertain each and every necessary contribution to an event (McGrath, 2005). Little is known about the factors that influence precisely *how many* candidates people consider in causal selection judgments from conjunctive and disjunctive structures. Furthermore, it remains to be seen whether the claims made by existing computational accounts of causal selection such as those proposed by Quillien (2020) or Morris et al. (2019) will generalize to structures involving more than simply two causal variables. The benchmark dataset provides a solution by including a third causal variable to these systems. This addition will allow us to explore the graded nature of responsibility attributions over a larger set of necessary or sufficient causal events.

The antecedent conditions for events in the real world are often far more complex than purely conjunctive or purely disjunctive systems can represent. Instead, many outcomes result from an intricate mosaic of these causal relationships (Gerstenberg et al., 2015). The benchmark dataset addresses this issue by probing responsibility judgments in contexts that combine different causal structures together in novel ways. More specifically, I characterize a mixed conjunctive system, in which a disjunctive causal structure is embedded within a conjunctive one such that two events must still both occur to bring about the token outcome, one of which, however, can occur through the disjunction of two distinct events. For example, consider a job opening, where applicants are required to have an advanced degree *and* coding proficiency in one language *or* another (e.g. Python or Javascript). The dataset also includes causal judgments in mixed disjunctive structures. In this case, a conjunctive causal structure is embedded within a disjunctive one such that two events are individually sufficient to bring about the token outcome, one of which, however, occurs through the conjunction of two distinct events. For example, consider a different job opening, where applicants are required to have an advanced degree *or* coding proficiency in two

programming languages (e.g. Python *and* Javascript). These novel composite structures allow for the generation of entirely new hypotheses for how causal selection unfolds in more complex environments.

As described above, our normative expectations interact with variable causal structures in interesting ways. Evidence for these effects has thus far come from work manipulating normality in broad, qualitative ways by describing events as either normal or abnormal. However, I believe normality can and should be construed in less binary ways. For the benchmark dataset, I parametrically modulate the influence of normality of causal events across a broad range of quantitative values. Covering such a vast space of possibilities will allow researchers to explore how causal judgments change in proportion to systematic changes in the normality of an event.

1.4.3 The current contribution

The purpose of the work reported in Chapter 4 of this dissertation is to discuss the impetus for a unified account of causal selection and introduce a large, benchmark dataset of human causal selection judgments I hope fosters new opportunities for researchers interested in the mechanisms of causal cognition.

The data in this benchmark are organized across four participant groups that differed in the causal structure they observed: i.) “Pure” 3-variable conjunction; ii.) “Pure” 3-variable disjunction; iii.) “Mixed” conjunction; iv.) “Mixed” disjunction. Importantly, the only difference in the stimuli presented to each group came from task instructions, which manipulated the causal structure through the description of which conditions were necessary and sufficient for the token outcome.

Participants made judgments about the causal contribution of three token events to a winning lottery outcome. More specifically, participants observed the token event of a red ball drawn from each of three jars containing red and blue balls. Across all groups, participants were made aware of the winning lottery outcome and the proportion of red and blue balls contained in each jar. The three token events that resulted in the winning lottery outcomes were held fixed such that, in the disjunctive case where at least one red ball was sufficient, participants observed a winning outcome resulting from three red balls being

drawn. This was done to match the causal events across conditions and observe only the effect of my manipulations of the causal structure and normality on causal selection. Finally, I varied the normativity of each causal event such that the probability of drawing a red ball from a given jar came from the set [.2, .4, .6, .8, 1], allowing us to systematically explore points across the continuous space of an event's probability from unlikely to certain. This large sample of human judgments contains 47,970 causal attribution ratings collected from 1,599 adult participants. It is my hope that this dataset may serve as a growing testbed for diverging theoretical models proposing to characterize how humans answer the question: Why?

Chapter 2.

2.1 Introduction

This paper asks whether there can be knowledge without belief. Or, more precisely, it asks whether people can represent someone else as knowing something without representing them as believing that thing. There are theoretical and empirical reasons to favor both a positive and negative answer to this question. We make progress on this question across five studies that offer a remarkably consistent answer using a combination of response time and fMRI data. Before turning to these results, it is worth explaining the motivation behind both the reasons for thinking there could not be knowledge without belief and the reasons for thinking there in fact could be.

2.1.1 Knowledge with belief

A standard way of thinking about the relationship between knowledge and belief across the cognitive sciences holds that knowledge is a relatively complicated mental state that may recruit or rely on more conceptually basic representations, like that of mere belief. It is not difficult to motivate such a perspective. For any case in which someone knows something, it seems intuitive enough that they must also at least believe that thing, since if they didn't even believe it, how could they know it? Perhaps then, attributing knowledge of something to others requires, at a minimum, representing them as believing that thing, and then additionally representing them as satisfying some additional criteria that are more specific to knowledge (e.g., that their belief is true, justified, formed through a reliable process, or what-have-you).

One sees aspects of this way understanding of the relationship between knowledge and belief both in standard philosophical analyses of knowledge (Ichikawa & Steup, 2016), and in recent

philosophically informed empirical work on knowledge attribution (Buckwalter et al., 2015; Dudley, 2018; Dudley et al., 2015; Rose & Schaffer, 2013).

Additionally, this understanding of the relationship between knowledge and belief sits well with the work in developmental and comparative psychology that has argued for a 'core' or 'innate' capacity for theory of mind that essentially involves an ability for belief representation (Kovács et al., 2010; Leslie et al., 2004; Onishi & Baillargeon, 2005; Stich, 2013). Researchers arguing for such a capacity have sought to offer evidence for an ability for false belief representation in non-human primates (Hayashi et al., 2020; Krupenye et al., 2016) and in human infants as young as 7 months (Buttelmann et al., 2009; Kovács et al., 2010; Onishi & Baillargeon, 2005). If this picture of the core capacity for theory of mind is correct, and essentially involves the ability to represent others' beliefs, then it would not be surprising if later-developing and more conceptually complicated mental state representations, e.g., 'imagining', 'being sure', or 'knowing' were built on top of conceptually primitive mental state representations, like mere belief, which are essential to the core capacity.

In short, this body of work collectively provides ample motivation for the view that ascriptions of knowledge may depend on or involve prior representations of belief.

2.1.2 Knowledge without belief

Against this view, an alternative picture that has gained attention across the cognitive sciences is that the capacity to represent knowledge is more basic than the capacity to represent belief, and thus representations of knowledge do not depend on representations of belief (Phillips et al., 2020). Once again, it is not particularly difficult to see how such a view might have intuitive merit: many mundane cases of reasoning about someone's mind seem to concern others' knowledge without concerning their beliefs. For example, if you missed the outcome of the last election and ask someone else if they know who won, you simply want to know what they know if they do know who won (and you do not care about their beliefs one way or the other).

More direct evidence for this second view comes from comparative research that has found that non-human primates often pass theory of mind tasks when they only involve representations of knowledge, but fail similar tests when they require belief representations (Horschler et al., 2020, 2019; Martin & Santos, 2014, 2016). Similarly, human infants are able to robustly pass theory of mind tasks that only require reasoning about others' knowledge long before they are able to pass tests that require belief representations (Luo & Johnson, 2009; Vouloumanos et al., 2014). Moreover, these findings concerning early-emerging knowledge representation have proven to be largely replicable, unlike the findings for early-emerging belief or implicit representations of belief (Holland & Phillips, 2020). Collectively, this research provides evidence for cases in which there is an ability to represent knowledge but no corresponding ability to represent belief. Obviously, in such cases, knowledge representation cannot depend on belief representation.

While these experimental tasks were designed to test for an underlying capacity for knowledge and belief representation, they do not explicitly employ the concepts of knowledge and belief. Importantly, however, other lines of research do explicitly employ these concepts and find a remarkably similar pattern. For example, across languages, young children begin demonstrating a linguistic competence for communicating about knowledge (and related factive mental states, such as 'seeing') before they demonstrate a linguistic competence for communicating about belief (or related non-factive mental states, e.g., what others 'think') (Bartsch & Wellman, 1995; Harris et al., 2017; Wellman et al., 2001). Additionally, research in experimental philosophy on the ordinary concept of knowledge has provided clear cross-cultural evidence that there are cases in which people are willing to attribute knowledge but unwilling to attribute belief (Myers-Schulz & Schwitzgebel, 2013; Yuan & Kim, 2021). And finally, recent work using EEG has revealed that answering questions about what someone believes requires greater cognitive resources than answering matched questions about what someone knows, as measured in terms of P3b amplitude (Bricker, 2020). Clearly, these cases of explicit representations of knowledge are unlikely to depend on prior (or simultaneous) representations of belief.

2.1.3 A core difference between knowledge and belief

Perhaps the most fundamental way in which knowledge and belief differ is that knowledge is *factive* while belief is not. That is, knowledge is an attitude that one can only represent others as having toward truths (R. Lakoff et al., 1973; Williamson, 2002), while one can believe both things that are true and things that are not (Nagel, 2017; Phillips & Norby, 2019). This difference in knowledge and belief is not hard to see: If you don't think it's true that Al Gore invented the internet, then you can't represent someone else as knowing that he invented the internet; you can, of course, represent them as believing it. And you can also represent them as believing any other arbitrary proposition, e.g., that Reagan killed all the birds in 1986 and replaced them with spies who work for the bourgeoisie. Representations of belief are not constrained by the truth; representations of knowledge are. Thus, an important difference between factive and non-factive attitudes, in general, is that factive attitude representations may be simpler because the content represented is necessarily consistent with your own, and thus does not have to be represented separately from your own understanding of the content (Phillips & Norby, 2019). Given that this difference holds for factive and non-factive attitudes in general, it would obviously hold for knowledge and belief in particular. We take advantage of this broader difference in trying to understand the relationship between knowledge and belief.

2.1.4 Present studies

Given the prior research, there is reason to think that it is possible that online knowledge ascriptions in human adults either may or may not depend on prior or simultaneous belief ascriptions. We contribute to this debate using a mix of response time studies and an analysis of fMRI data while human adults are making ascriptions of both knowledge and belief. In Experiment 1, we find that people can accurately evaluate others' knowledge before they evaluate their beliefs. In Experiment 2, we rule out the possibility that this response-time pattern is explained by pragmatic differences, both by measuring the felicity of knowledge and belief ascriptions and by collecting data from high-functioning individuals with Autism Spectrum Disorder, who are known to differ in their pragmatic inferences. We find that, similar to

neurotypicals, this cohort also attributed knowledge to agents faster than they attributed beliefs, and that response-time differences are unlikely to be explained by pragmatics. In Experiment 3, we demonstrate that the relative speed advantage of knowledge over belief ascriptions occurs cross-linguistically and is not accounted for by differences in word frequency. In Experiment 4, we find that this response-time difference generalizes to a larger class of factive and non-factive attitudes (to which knowledge and belief respectively belong). And in Experiment 5, we show that the neural response pattern that occurs when making evaluations of others' beliefs is absent when making similar evaluations of knowledge. Together, these studies demonstrate that human adults can attribute or deny knowledge states without prior or simultaneous evaluations of belief states. These findings collectively support the view that knowledge representations are a basic and distinct way in which we understand others' minds.

2.2 Experiment 1a-b: Knowledge before Belief

A simple way of determining whether people can make evaluations of knowledge in the absence of evaluations of belief is to investigate the speed with which these evaluations are made. If participants' evaluations of knowledge require or involve evaluations of belief, we would expect evaluations of knowledge ascriptions to be slower than evaluations of similar belief ascriptions. We began with a simple task that investigated the response times for evaluations of knowledge and belief ascriptions. Participants read about various people in a variety of different scenarios and then made truth value judgments of knowledge and belief ascriptions to those agents. In some cases, the knowledge and belief ascriptions were true, in others, they were false. We then compared the speed with which participants were able to correctly evaluate these mental state ascriptions.

2.2.1 Analysis approach

In Experiments 1-4, response times for trials on which participants correctly assessed the truth of the knowledge and belief statements were analyzed with linear mixed-effects models using the lme4 package in R (Bates et al. 2014), with both participants and scenarios included as random factors. Participants were excluded from the analysis if they answered less than 67% of the questions correctly or if their *mean* response time was less than 1000*ms* or greater than 4000*ms*. We additionally excluded all trials on which the response was given in less than 1000*ms* or longer than 4500*ms*. We applied these same criteria to all of the response-time analyses in all of the experiments reported.

2.2.2 Participants

Two hundred participants (M_{age} = 32.76, SD_{age} = 12.67; 108 females) were recruited through a psychology-based website (http://www.moralsensetest.com). Experiment 1b was an exact replication with 501 participants recruited through Amazon Mechanical Turk.

2.2.3 Stimuli and procedure

Participants began by completing a demographic questionnaire, and two practice trials in which they were familiarized with the task they would be completing. Participants then completed twenty-four trials which involved reading a short vignette about an agent and then deciding whether a sentence about the story was true or false. Participants were instructed to indicate their responses as quickly as possible by pressing one of two keys on their keyboard. Twelve of these trials were 'distractor' trials in which participants were asked simple comprehension questions that did not mention the agent's mental states. These were included to prevent participants from anticipating the sentences they would be evaluating. In the remaining twelve trials, participants read a vignette that described the agent as either having true information about some proposition *p* (as in **True Information**), simply being ignorant of *p* (as in **No Information**), or believing some proposition q that was both false and inconsistent with p (as in **False Information**).

True information: Mira looks at the night sky with her telescope. She owns the most accurate books on the locations of the different planets throughout the year. Mira reads in her astronomy books that she can see Neptune through her telescope, and she waits until it's dark enough outside. She points her telescope toward the coordinates that her books specify for Neptune, and sees a bright dot in the middle of the sky. That bright dot is Neptune. She is excited that she found the planet she was looking for so easily.

No Information: Mira likes looking at the night sky with her telescope. She owns the most accurate books on the locations of the different planets throughout the year. It is night and Mira decides not to read her astronomy books and instead just look through her telescope. Ignoring her book, she sets up her telescope and points it toward a group of dots that catch her attention. She looks into the telescope and she sees a bright dot in the middle of the sky. That bright dot is actually Neptune.

False information: Mira likes looking at the night sky with her telescope. She owns the most accurate books on the locations of the different planets throughout the year. It is night and Mira reads in her astronomy books that she can see Mercury through her telescope. Misreading her book, she sets up her telescope and points it toward the coordinates that her books specify for Neptune, which also happens to be in the sky. She looks into the telescope and she sees a bright dot in the middle of the sky. That bright dot is actually Neptune.

On each of the twelve test trials, participants were asked to evaluate the truth or falsity of a sentence about either the agent's knowledge, as in **Know**, or belief, as in **Think**. Critically, these sentences always concerned the proposition that was true:

Know: Mira knows she is looking at Neptune.

Think: Mira thinks she is looking at Neptune.[1]

Scenarios in all three conditions were constructed in such a way that the answer to both questions was true (i.e., True Information) or the answer to both questions was false (i.e., No Information and False Information), which allows us to directly compare response times across the two ascription types (Know vs. Think). We used Latin-square randomization to ensure that each participant saw all 12 distractor scenarios and all 12 test scenarios in random order, and judged two 'knows' ascriptions and two 'thinks' ascriptions for each of the three Information Conditions ('True Information' vs. 'No Information' vs. 'False Information').

2.2.4 Experiment 1a-b Results

Experiment 1a Results

The overall analysis of participants' response times revealed no main effect of Information Condition, $\chi2(2) = 1.565$, $p = 0.457$, and no Information Condition × Ascription Type interaction, $\chi2(2) = 1.602$, $p = 0.4492$. However, there was a significant main effect of Ascription Type, $\chi2(1) = 20.057$, $p < 0.001$), such that participants were faster to correctly assess the truth of statements about whether the agent knows something (M_{rt} = 2550.46ms, SD_{rt} = 739.06) than statements about whether the agent thinks something (M_{rt} = 2655.43ms, SD_{rt} = 751.27).

---

[1] We used 'thinks' instead of 'believes' in these studies to better equate for word frequency and length.

[2] The fixed and random effects structure for the full model was specified as: *response.time ~ info.condition ∗ ascription.type + (1|scenario) + (info.condition ∗ ascription.type|subj).* We were not able to include random slopes for the scenario because the crossed nature of the random effects in our experiment prevented convergence. We employ a similar analysis approach throughout, except where explicitly noted.

Experiment 1b Results

Experiment 1b was a direct well-powered replication of Experiment 1a, with the only difference being that participant recruitment was through Amazon's Mechanical Turk. The analysis of participants' response times again revealed no main effect of Information Condition, $\chi^2(2) = 4.126$, $p = 0.127$, but did reveal a main effect of Ascription Type, ($\chi^2(1) = 27.112$, $p < 0.001$), such that participants were faster to correctly assess the truth of statements about whether the agent knows something (M_{rt} = 2459.37ms, SD_{rt} = 739.02) than statements about whether the agent thinks something (M_{rt} = 2517.93ms, SD_{rt} = 731.17). Additionally, there was a significant Information Condition × Ascription Type interaction, $\chi^2(2) = 12.851$, $p = 0.002$. Given the identical methods of Study 1a and 1b, we next combined the data from both experiments to ask whether the interaction was significant across all of the data, and to provide the best estimate of the other effects.

Combined Analyses

The combined analysis continued to reveal no main effect of Information Condition, $\chi^2(2) = 2.68$, $p = 0.262$, but did reveal a main effect of Ascription Type, ($\chi^2(1) = 47.013$, $p < 0.001$), such that participants were faster to correctly assess the truth of statements about whether the agent knows something (M_{rt} = 2485.33ms, SD_{rt} = 740.05) than statements about whether the agent thinks something (M_{rt} = 2556.11ms, SD_{rt} = 740.05). Moreover, we continued to observe a significant Information Condition × Ascription Type interaction, $\chi^2(2) = 12.385$, $p = 0.002$.

Planned pairwise comparisons were carried out using the Estimated Marginal Means package in R (Lenth et al., 2022). These analyses revealed that the largest difference in response times was found in the No Information conditions, $t(477) = -6.839$ $p < 0.0001$, followed by a somewhat smaller difference in the True Information conditions, $t(473) = -3.421$ $p = 0.009$. The smallest response time difference was found in the False Information conditions, $t(476) = -2.287$, $p = 0.201$. Response time differences in each

information condition showed the same pattern such that participants were faster to evaluate knowledge ascriptions than belief ascriptions (see Figure 2.1).

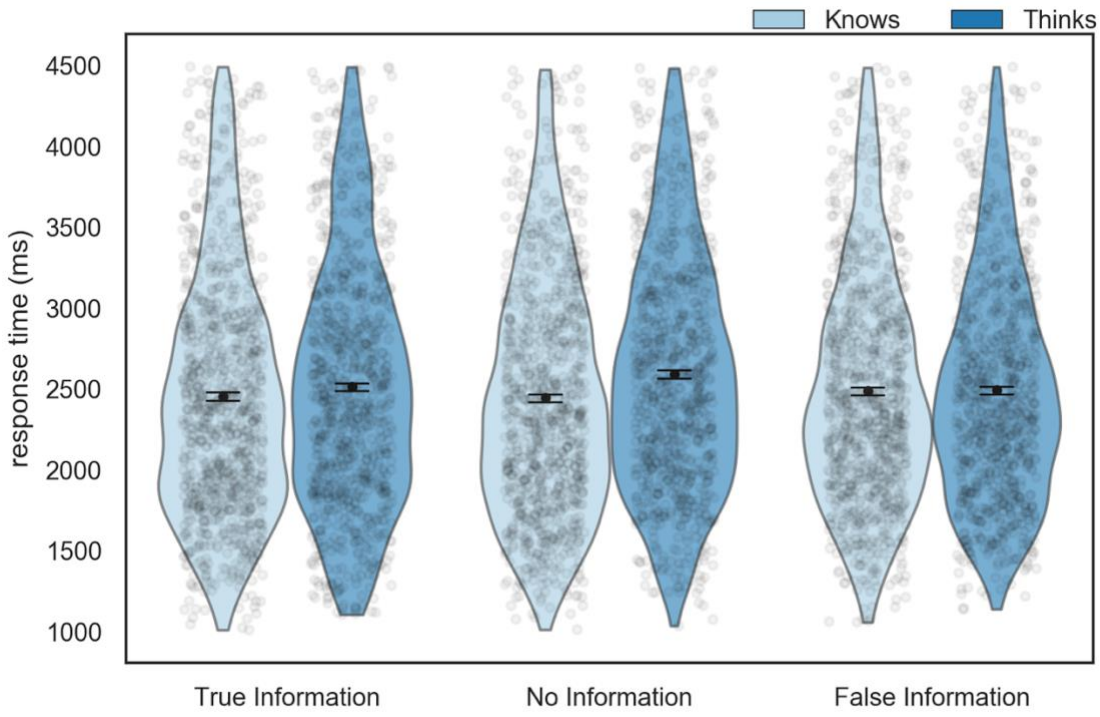


Figure 2.1, Response times for correct evaluations of knowledge and belief ascriptions (dark plots) as a function of Information Condition. Error bars depict +/- 1 SEM.

2.2.5 Discussion

If evaluations of others' knowledge involve or require evaluations of belief, then we would expect that correctly determining whether someone knows something would take at least as long as determining whether someone believes that thing. We tested this prediction but found clear evidence that the opposite is true: correct evaluations of knowledge are made *before* correct evaluations of belief. In other words, the evidence suggests that participants are both attributing knowledge (in the True Information condition) and denying knowledge (in the No Information conditions) before they have been

able to determine whether the agent believes the relevant claim. A further question is why we observed an interaction effect, such that the difference in response times in assessing knowledge and belief ascriptions differed in the various information conditions. Before making much of this unpredicted interaction effect, we first want to ask whether it replicates, and ensure that the observed effects are not arising from some simpler, alternative explanation.

One potential concern with this initial study is that it involves cases in which some of the belief ascriptions may be pragmatically odd. Specifically, in cases where an agent meets all of the criteria for knowledge (as in the True Information example), it may sound unnatural to describe the agent as believing some proposition (as in **Think**) rather than describing the agent as knowing that proposition (as in **Know**)[3]. After all, the agent does not merely think the proposition in question, but also knows it. In such cases, while belief ascriptions will, of course, be strictly true, they may also be pragmatically odd or infelicitous. We next investigate whether this kind of pragmatic effect can explain the pattern of response times we observed.

2.3 Experiment 2: Linguistic Pragmatics & ASD

In this Experiment, we aimed to test for the impact of pragmatic infelicity while also asking whether the observed effect generalizes to participants with Autism Spectrum Disorder (ASD). To ask whether differences in the pragmatic understanding of the sentences participants evaluated might explain the observed response time difference between knowledge and belief ascriptions, we measured the extent to which the sentences participants evaluated were perceived as pragmatically odd by participants. While pragmatic theories do not strictly predict a difference in the felicity of knowledge and belief ascriptions

[3] This kind of pragmatic effect would be predicted on a number of different theories, e.g., (Heim, 1991; Hirschberg, 1985).

when the agent is ignorant or has a false belief (as both ascriptions in these cases are simply false), we decided to collect felicity judgments for all of the sentences used in the previous experiment.

Additionally, we also investigated whether the previously observed response time effect generalizes to a sample of participants with ASD. Prior research has demonstrated that individuals with autism present impairments in understanding pragmatic implicatures, especially when they involve reasoning about others' mental states (Kissine, 2012). Additionally, a separate well-established line of research on autism has demonstrated impaired mentalizing abilities as assessed through classic false-belief paradigms (Baron-Cohen et al., 1985). Accordingly, participants with ASD make an ideal population for exploring the previously observed response time difference: If pragmatic differences are responsible for the comparative delay observed in correctly evaluating belief ascriptions, we should expect differences in pragmatic understanding to be reflected in the response time patterns between the ASD group and the neurotypical control group. Alternatively, if the pattern of response times found in Experiment 1 instead resulted from knowledge assessments being made in the absence of calculations of others beliefs, we expect participants with autism to demonstrate the same, or more extreme differences in response times between knowledge and belief ascriptions, as they may specifically have difficulty with belief representation but not knowledge representations (Deschrijver & Palmer, 2020; Phillips et al., 2020).

2.3.1 Participants

Inclusion criteria required participants to be adults, fluent in English, and complete the experiment on a personal computer or laptop with a standard keyboard. The 611 participants included in subsequent analyses each had of a mean accuracy ≥ 67% and a mean response time > 1000ms and < 4000ms. 389 Participants were recruited from Prolific (app.prolific.co) to the neurotypical group, (M_{age} = 37.57, SD_{age} = 13.12, 50.9% female, 45% male, 3.34% other, 0.77% not disclosed).

Crowdsourcing marketplaces such as Amazon's Mechanical Turk or Prolific do not provide stratified participant pools based on an autism diagnosis. Therefore, recruitment of participants in this

group came from study advertisements posted in various online forums and websites of organizations dedicated to the autism community during the period from August-October 2022. Given this open recruitment strategy, advanced security measures were taken to ensure that data collected from this group were provided by earnest participants[4] (Dennis et al., 2019). A short screener required participants to endorse eligibility criteria, including fluency in English, being at least 18 years of age, a computer or laptop to complete the experiment (no phones/tablets, etc.), and self-report of a clinical autism diagnosis. Overall, 222 participants recruited to the autism group met the same accuracy and mean response time criteria described above for neurotypicals, (M_{age} = 28.12, SD_{age} = 5.15, 43.24% female, 51% male, 4.95% other, 0.45% not disclosed). Importantly, participants in the autism group scored significantly higher on the 10-item Autism Quotient scale ($M_{AQ\text{-}10}$ = 5.74, $SD_{AQ\text{-}10}$ = 2.7) than participants in the neurotypical group ($M_{AQ\text{-}10}$ = 3.72, $SD_{AQ\text{-}10}$ = 2.26), $t(397.7) = -9.43$, $p < 0.0001$.

2.3.2 Stimuli and procedure

The experiment was conducted in two blocks. The first consisted of the same stimuli and procedures as Experiment 1a-b, in which the 3 (Information Condition: No Information, True Information, False Information) x 2 (Ascription Type: Know, Think) design conditions were randomized and counterbalanced across 12 unique vignette contexts. 12 distractor trials, in which participants were asked to evaluate the veracity of simple facts about a vignette, were also presented, but excluded from all analyses. Response times were recorded while participants evaluated mental state ascriptions as true or

[4] After screening, eligible participants received an automated email invitation to participate containing a unique study access link that could only be used once. The experiment was served to eligible participants from a custom-built encrypted server (https://www.linode.com/), secured with a firewall and other security precautions. Connections to the experiment server were tested to ensure that the experiment was not accessed through a virtual privacy network, proxy, relay network, or tor node. Experiment access was also restricted to connections from English-speaking countries according to ISO-639 standards. Finally, two unrelated open-ended free text response questions were required items in the experiment and manually screened for nonsensical or suspicious responses.

false. The second block consisted of the same conditions, such that the 12 vignettes and mental state ascriptions in the first block were presented a second time in randomized order. Each participant completed a training session in which the felicity judgment task was thoroughly explained. In addition, they completed two practice trials using statements that were clearly felicitous or clearly infelicitous, and were given feedback on their responses. On the test trials, participants were asked to judge how much the target statement ascribing a mental state to the agent in the vignette seemed infelicitous, and responded on a Likert scale from 1("sounds very normal") to 7 ("sounds very weird"). Following completion of the first two blocks, participants completed a 10-item Autism Quotient scale, provided demographic information, and were debriefed.

2.3.3 Results

Infelicity ratings

Focusing first on our measure of the pragmatic oddness of the sentences used, we tested for a three way interaction between Participant Group, Information Condition, and Ascription Type. We found no significant three-way interaction $\chi^2(2) = 0.0591$, $p = .970$. However, there were clear differences in infelicity judgments between autism and neurotypical groups. Neurotypical participants seem to better grasp that false statements may still be felicitous, whereas participants in the autism group judged false mental state ascriptions of both knowledge and beliefs as relatively infelicitous (Figure 2, A). This pattern is evidenced by a significant group by agent state interaction effect on infelicity judgments $\chi^2(2) = 175.55$, $p < 0.0001$. When agents are described as having no information (*Figure 2.2A-B*, center plots), participants in the autism group rated mental state ascriptions (of both knowledge and beliefs) more infelicitous ($M_{\text{infelicity}} = 4.52$, $SD_{\text{infelicity}} = 2.62$) than participants in the neurotypical group ($M_{\text{infelicity}} = 2.41$, $SD_{\text{infelicity}} = 2.03$), $t(620.38) = 14.33$, $p < 0.0001$. Similarly, when agents are described as having false information (*Figure 2.2A-B*, right-most plots), participants in the autism group, again, judged mental state ascriptions more infelicitous ($M_{\text{infelicity}} = 4.86$, $SD_{\text{infelicity}} = 2.64$) than neurotypical participants ($M_{\text{infelicity}} = 2.32$, $SD_{\text{infelicity}} = 2$),

$t(620.02) = 16.4$, $p < 0.0001$. There was no group difference, however, in infelicity judgments for agents described as having true information, $t(623.42) = -1.22$, $p = 0.83$.

While, we found no Participant Group by Ascription Type interaction effect on infelicity judgments, $\chi^2(2) = 0.266$, $p = 0.606$, we did observe a significant interaction of Information Condition x Ascription Type on infelicity judgements, $\chi^2(2) = 25.398$, $p < 0.0001$, which was in line with the predictions of pragmatic theories (Heim, 1991; Hirschberg, 1985). Across both autism and neurotypical groups, belief ascriptions ($M_{\text{infelicity}} = 1.88$, $SD_{\text{infelicity}} = 1.62$) were judged significantly more infelicitous than knowledge ascriptions ($M_{\text{infelicity}} = 1.58$, $SD_{\text{infelicity}} = 1.36$) for contexts in which agents have true information, $t(65.7) = -4.29$ $p < 0.0001$. This pattern disappears when agents are described as having false information, $t(71.15) = 1.86$, $p = .432$, or no information, $t(81.14) = 1.32$, $p = .774$. There was no overall main effect of ascription type on infelicity judgments, $\chi^2(2) = 0.576$, $p = 0.448$. Critically, we can also use these infelicity judgments to control for possible effects of infelicity in subsequent response time analyses because we collected these data for each mental state ascription within subjects.

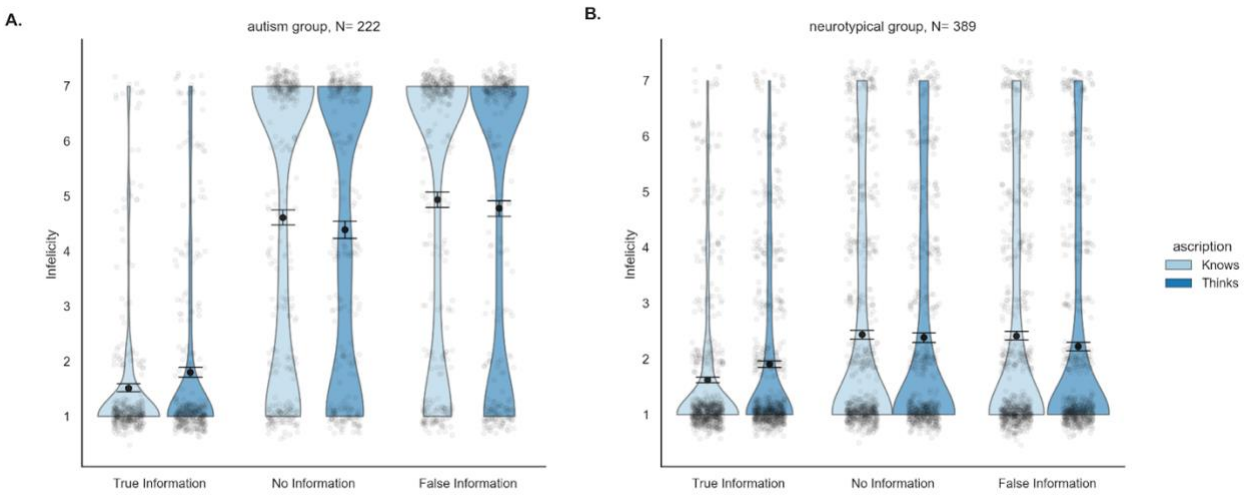


*Figure 2.2, Mean infelicity ratings from **A**. autism group, **B**. neurotypical group for knowledge ascriptions (light bars) and belief ascriptions (dark bars) as a function of Information Condition. Error bars depict +/- 1 SEM.*

Response Times

Controlling for perceived infelicity, we observed a marginally significant three-way interaction of Participant Group, Information Condition, and Ascription Type on response times, $\chi^2(2) = 5.719$, $p = 0.057$ (*Figure 2.3 A-B*). This three-way interaction can be explained by examining Information Condition x Ascription Type interactions within each participant group separately. Participants in the autism group did not show an Information Condition x Ascription Type interaction effect, $\chi^2(2) = 0.234$, $p = 0.89$. In the neurotypical group, however, we found a significant interaction effect of Information Condition x Ascription Type on response times, $\chi^2(2) = 28.742$, $p < .001$. Pairwise tests demonstrated that this was driven by the No Information condition, where knowledge ascriptions (M_{rt} = 2242.87ms, SD_{rt} = 716.41) were evaluated significantly faster than belief ascriptions (M_{rt} = 2511.42ms, SD_{rt} = 762.56), $t(35.83) = -7.29$, $p < 0.0001$ (Figure 3B, middle). Similar comparisons did not reveal a significant difference in response times between knowledge and belief ascriptions in the True Information ($t(31.57) = -2.17$, $p = .278$) or False Information ($t(32.43) = -2.51$, $p = .15$) conditions, though the effect was always in the same direction.

There was also a significant Participant Group x Information Condition interaction effect on response times, $\chi^2(2) = 9.804$, $p = 0.007$. Post hoc tests of this interaction were computed using the Estimated Marginal Means package in R, and revealed that participants in the autism group (M_{rt} = 2213.21, SD_{rt} = 785.51) were significantly faster than neurotypicals (M_{rt} = 2386.87, SD_{rt} = 717.5) at ascribing either type of mental state to agents with true information, $t(623.08) = -4.36$, $p = 0.0001$ (*Figure 2.3A*, left vs *2.3B*, left). The same pairwise comparisons did not reveal group differences in overall response time in the False Information ($t(682.56) = -1.73$, $p = .509$) and No Information ($t(657.31) = -1.69$, $p = .537$) conditions, though here too ASD participants were still somewhat faster in responding.

Critically, there was no interaction between Participant Group and Ascription Type on response time, $\chi^2(2) = 0.135$, $p = 0.713$, indicating that the significant main effect of Ascription Type on response times exists for both neurotypicals and autistic participants equally. Replicating the prior experiments, participants in the neurotypical group were significantly faster at evaluating the truth of knowledge

ascriptions (M_{rt} = 2295.04ms, SD_{rt} = 693.78) than they were at evaluating the truth of belief ascriptions (M_{rt} = 2429.38ms, SD_{rt} = 724.65), $t(17.07) = -5.79$, $p < 0.0001$ (*Figure 2.3B,* light vs. dark). Similarly, we found that participants in the autism group were also significantly faster at evaluating the truth of knowledge ascriptions (M_{rt} = 2174.51ms, SD_{rt} = 771.54) than they were at evaluating the truth of belief ascriptions (M_{rt} = 2321.68ms, SD_{rt} = 823.27), $t(39.62) = -4.36$, $p < 0.0001$ (Figure 3A, light vs. dark).

To further test whether the differences we found in response times for evaluations of belief and knowledge ascriptions could be explained by pragmatic infelicity, we examined the relationship between response times and infelicity judgments directly. We found a significant correlation between infelicity judgments and response times in the neurotypical group such that mental state ascriptions judged as more infelicitous took longer to evaluate, $r = .056$, $p = .0004$ (*Figure 2.3D*). This relationship was not found in the autism group $r = -.04$, $p = .101$ (*Figure 2.3C*).

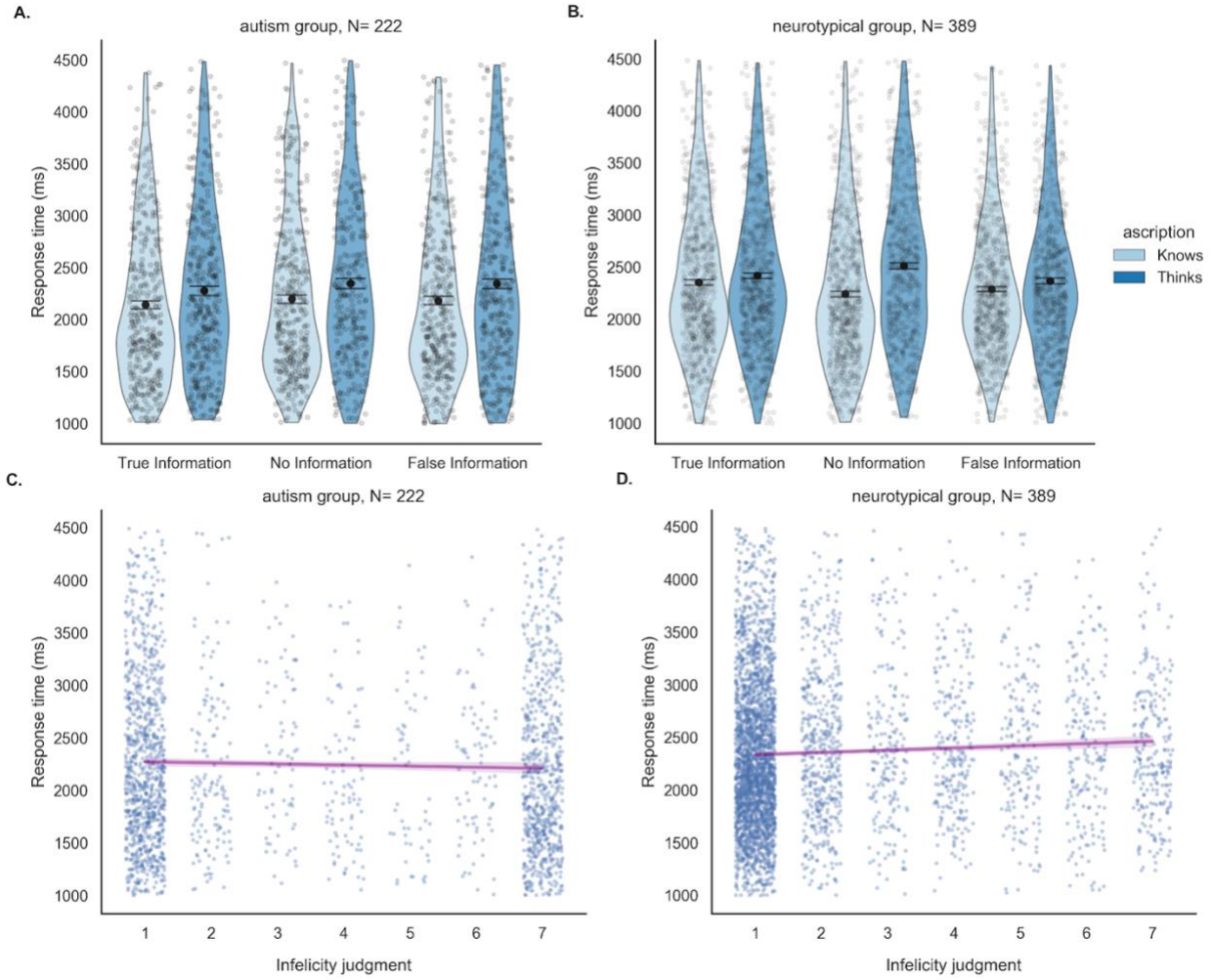


Figure 2.3, ***A-B.*** *Response times of autism (**A**) and neurotypical (**B**) groups for correct evaluations of knowledge ascriptions (light plots) and belief ascriptions (dark plots) as a function of Information Condition. Error bars depict +/- 1 SEM.* ***C-D****. Response times by ascription infelicity judgments for autism (**C**) and neurotypical (**D**)*

Autism Quotient questionnaire and response times

Both groups of participants completed the 10-item psychometric Autism Quotient scale (AQ-10) to assess traits associated with the autism spectrum. Persons scoring 6 or above on the AQ-10 are recommended for a more comprehensive assessment for autism (Allison et al., 2012).

To examine a more continuous relationship between autism symptomatology and response times in correctly assessing mental state ascriptions, we computed the difference between each subject's average response time for belief and knowledge ascriptions, such that more positive values indicate that knowledge ascriptions were faster than belief ascriptions while more negative values indicate that belief ascriptions were faster. We then asked whether self-reported autism symptoms, as measured by the AQ-10, predict the difference in speed with which knowledge and belief ascriptions are correctly evaluated. We did not find a significant relationship, though to the extent there is a systematic pattern, higher AQ-10 scores are associated with a *bigger* difference in response times, $F(1, 609) = 1.12$, $p = 0.29$ (Figure 4).

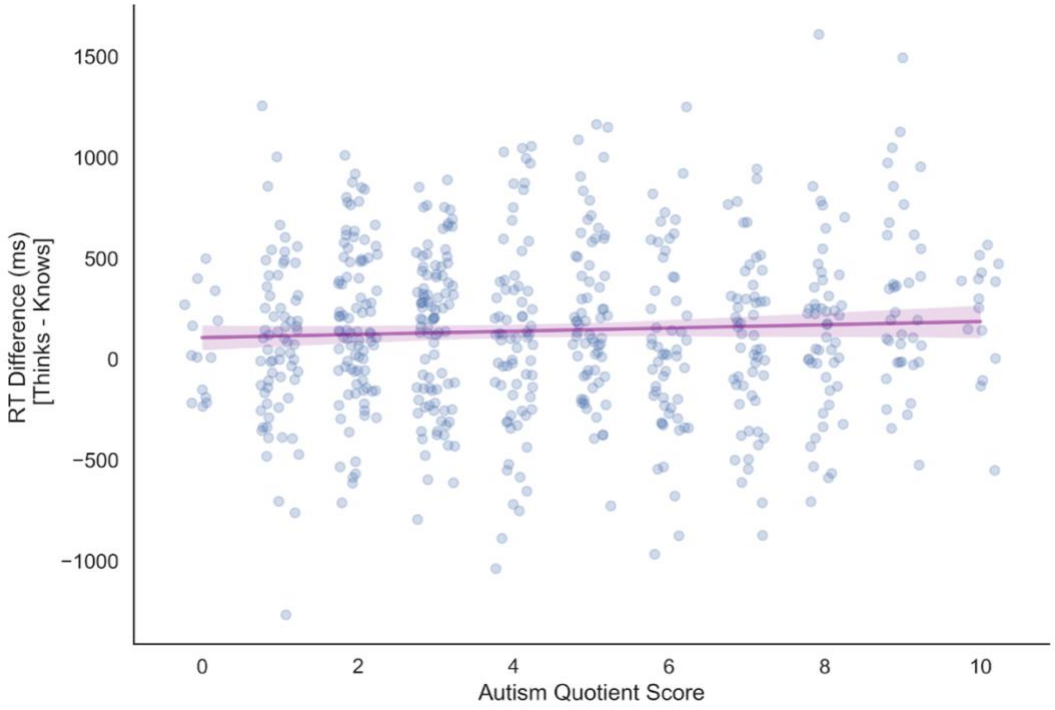


Figure 2.4, Difference in participant mean response time between belief and knowledge attribution trials as a function of participants' score on the Autism Quotient - 10 scale across all participants

2.3.4 Discussion

In order to rule out the possibility that response time differences between knowledge and belief ascriptions did not result from differences in the pragmatics of the two kinds of ascriptions, we asked a new group of participants to evaluate how felicitous these sentences were after first quickly judging the truth of the sentences. We also explored whether or not the observed pattern of response time differences extends to high-functioning individuals with autism, despite observed differences in tendency and ability in attributing mental states to others. We found that mental state ascriptions to agents with no information and agents with false information were judged more infelicitous by autistic participants, suggesting that autistic subjects seem to find false statements pragmatically odd. Moreover, as predicted by theories of pragmatics, belief ascriptions to agents meeting the criteria for knowledge were judged significantly more infelicitous than knowledge ascriptions for both the neurotypical participants and participants with ASD. Controlling for any perceived infelicity, we replicated the key response time patterns observed in Experiment 1. Most importantly, we found that both autistic and neurotypical participants were able to correctly evaluate knowledge ascriptions before they were able to correctly evaluate belief ascriptions. Furthermore, the difference in response times between knowledge and belief ascriptions was also not significantly modulated by autism symptoms, as assessed by the AQ-10. Taken together, these results suggest that the response time patterns found in Experiments 1 and 2 do not arise from pragmatic differences in knowledge vs. belief ascriptions, and are similarly robust across samples from neurotypical populations and populations with ASD.

2.4 Experiment 3: Lexical Frequency & French Language

We next asked whether the observed response time difference between 'know' and 'think' generalized to languages other than English. To do this, a highly similar experiment was conducted in French using the mental state verbs 'savoir' and 'penser' instead of 'know' and 'think' (respectively). In addition to providing a general test of whether the observed effect could have arisen from idiosyncratic features of the English used in the prior experiment, French provides a particularly strong test because, unlike English, the French term used for belief ascriptions 'penser que' is ≈ 1.49 times more frequent than French term used for knowledge ascriptions 'savoir que.'[5] Thus, if the difference in response times merely reflected lexical frequency, we should expect evaluations of beliefs to be faster than evaluations of knowledge in French, which is the opposite of the prediction of our more general theory of mind based account.

2.4.1 Participants

150 participants (M_{age} = 37.7, SD_{age} = 12.16; 83 females) were recruited and paid €0.75 through Foule Factory (https://www.foulef actory.com/).

2.4.2 Stimuli and procedure

The methods and procedures in this experiment were similar to that of Experiment 1a-b, except that the study was translated into French, and the English names were replaced with more standard French names. Thus, for example, rather than the **Know** and **Think** example sentences in Experiment 1, participants evaluated **Savoir** and **Penser** as below.

---

[5] Lexical frequency was computed using Google NGram using data from 2007-2008, which was the most recent year available at the time of calculation. For both French and English, we summed the frequency of the following forms of 'Savior' and 'Penser' or 'Know' and 'Think': infinitival, first person singular, second person singular, third person masculine, third person feminine, first person plural, second-person plural (French only), third-person plural (English), and masuline and feminine third person plural (French). We then divided the summed frequency of 'know' or 'savoir' by the summed frequency of 'think' or 'penser' respectively.

Savoir: Nora sait qu'elle regarde Neptune.

Penser: Nora pense qu'elle regarde Neptune.

2.4.3 Results

As in Experiment 1 and 2, data were excluded at the participant- and trial-level, and then analyzed using linear mixed-effects models. We found no main effect of Information Condition, $\chi^2(2) = 4.022$, $p = 0.134$. However, there was again a highly significant main effect of Ascription Type, $\chi^2(2) = 22.246$, $p < 0.001$, such that participants were faster to correctly assess the truth of statements about what the agent knows (M_{rt} = 2546.8ms, SD_{rt} = 719.03) than statements about what the agent thinks (M_{rt} = 2701.13ms, SD_{rt} = 716.9). Once again, there was also a significant Information Condition × Ascription Type interaction, $\chi^2(2) = 7.434$, $p = 0.024$ (see Figure 5).

Planned pairwise comparisons revealed that participants' response times only differed significantly in No Information, $t(89) = 5.091$, $p < 0.001$, $d = 0.567$ and True Information, $t(95) = 1.981$, $p = 0.05$, $d = 0.198$, conditions, but no significant effect was found in the False Information condition, $t(89) = 1$, $p = 0.32$, $d = 0.114$.

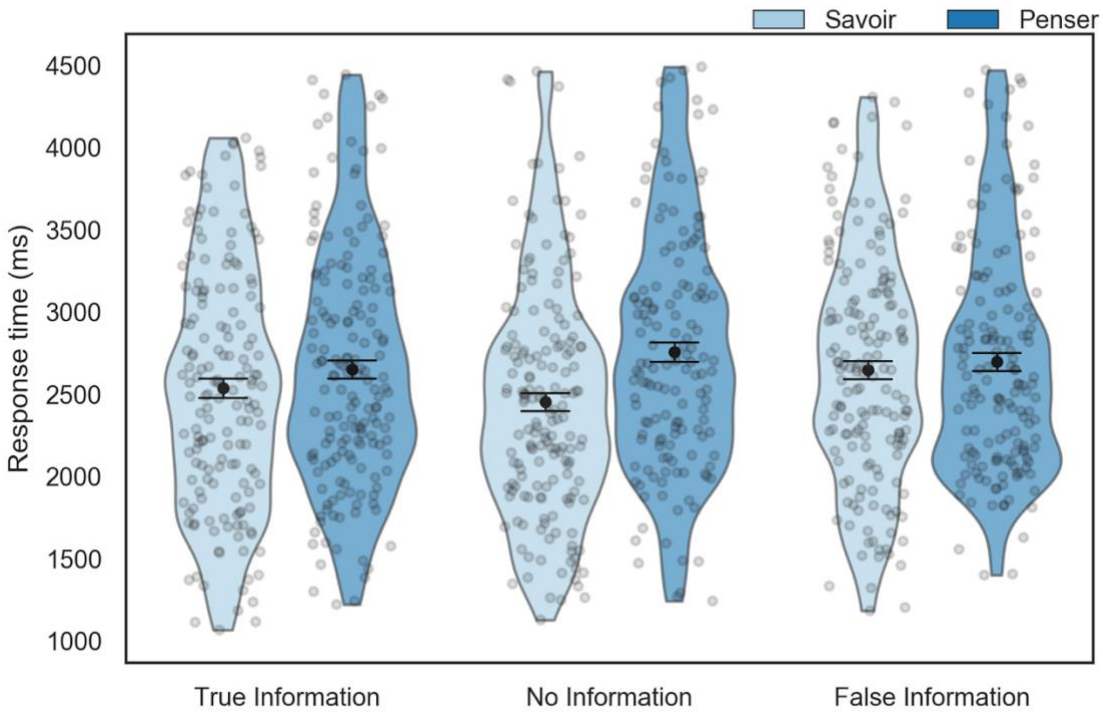


Figure 2.5, Response times for correct evaluations of knowledge ("Savoir") ascriptions (light plots) and belief ("Penser") ascriptions (dark plots) as a function of Information Condition. Error bars depict +/- 1 SEM.

2.4.4 Discussion

This third experiment provides cross-linguistic evidence that evaluations of knowledge ascriptions occur prior to equivalent belief ascriptions, and thus that evaluations of others' knowledge are made in the absence of similar evaluations of their beliefs. In addition, it provides clear evidence against an explanation of this pattern in terms of lexical frequency.

An important remaining question is whether the observed difference is specific to knowledge and belief or whether it reflects a more general difference between different categories of mental state verbs of which knowledge and belief are a part. We pursue this possibility in Experiment 4.

2.5 Experiment 4: Factive & Non-Factive States

Knowledge and belief are just two representative examples of the broader classes of factive and non-factive mental state attitudes. Similar to knowledge, you cannot represent someone as 'seeing' or 'hearing' something that is false, because these attitudes are factive. And similar to belief, you can represent someone as 'assuming' or 'guessing' things that are false, because these attitudes are non-factive. Accordingly, one possibility is that people will be generally faster to evaluate factive attitude ascriptions than non-factive attitude ascriptions. A plausible reason to predict this general difference is that factive mental state representations may be simpler because the content represented is necessarily consistent with one's own understanding, and thus does not have to be represented separately (Phillips & Norby, 2019). If this prediction proves true, it could point toward a more general explanation of why people are faster to evaluate knowledge ascriptions than belief ascriptions.

To clarify, our prediction is not that every factive attitude will be evaluated faster than every non-factive attitude; after all, there are many factors that will jointly determine evaluation time. Our prediction is instead that–other things being equal–it will take longer to evaluate the truth of a given non-factive attitude than an otherwise similar factive attitude. We next investigate this possibility by considering the speed with which participants evaluate a larger range of factive and non-factive mental state ascriptions.

2.5.1 Participants

250 participants (M_{age} = 33.43, SD_{age} = 9.32; 126 females) were recruited and paid through Amazon Mechanical Turk.

2.5.2 Stimuli and procedure

The methods and procedures in this experiment were similar to the preceding studies except that the term 'know' in the mental state ascription was replaced by one of a set of factive attitude verbs ("realize", "is aware", "understand", "recognize"), and the term "think" was replaced by a set of non-

factive attitude verbs (“believe”, “guess”, “assume”, “imagine”). Thus, for example, instead of evaluating the truth or falsity of **knows** or **thinks** as Experiment 1-2, participants may have evaluated the truth or falsity of **factive** or **non-factive**, respectively.

Factive: Mira recognizes that she is looking at Neptune.

Non-factive: Mira believes that she is looking at Neptune.

Critically, these factive and non-factive terms were chosen such that the non-factive terms were both shorter in length and more frequent in use than the factive terms (*Table 2.1*), so that the predicted difference in response times (factive < non-factive) could not arise from differences in word length or lexical frequency. Because each specific factive and non-factive verb has unique features of their meaning, ascriptions with each mental state verb did not always make sense in the context of the twelve different scenarios. Accordingly, for each scenario, we chose a pair of factive and non-factive terms that each made clear conceptual sense in the context described. Each factive and non-factive term was used in three scenarios.

Mental State	Factivity	Length	Frequency
“realize”	factive	7	.00338
“recognize”	factive	9	.00395
“understand”	factive	9	.01518
“is aware”	factive	8	.00610
“believe”	non-factive	7	.01716
“guess”	non-factive	5	.00339
“imagine”	non-factive	7	.00393
“assume”	non-factive	6	.00414

Table 2.1, Mental state verbs, their factive or non-factive status, word length and lexical frequency used in mental state attitude ascription statements for Experiment 4.

2.5.3 Results

As in the previous experiments, data were excluded at the participant- and trial-level, and then analyzed using an identical set of linear mixed-effects models, only now the effects calculated for Ascription Type reflect whether a verb is factive vs non-factive rather than the verb being ‘know’ vs. ‘think’[6]. This revealed a small main effect of Information Condition, $\chi^2(2) = 7.143$, $p = 0.028$, but no

[6] The fixed and random effects structure for the full model was specified as: *response.time ~ info.condition ∗ factivity.type + (1|scenario) + (info.condition ∗ factivity.type|subj)*. We were not able to include random slopes for the scenario because the crossed nature of the random effects in our experiment prevented convergence.

Information Condition × Ascription Type interaction, $\chi^2(2) = 0.686$, $p = 0.709$. More importantly, we again observed a significant main effect of Ascription Type, $\chi^2(2) = 6.39$, $p = 0.011$, such that participants were faster to correctly assess the truth of ascriptions involving factive attitudes (M_{rt} = 2361.18ms, SD_{rt} = 694.73) than ascriptions involving non-factive attitudes (M_{rt} = 2420.31ms, SD_{rt} = 709.58) (see Figure 6).

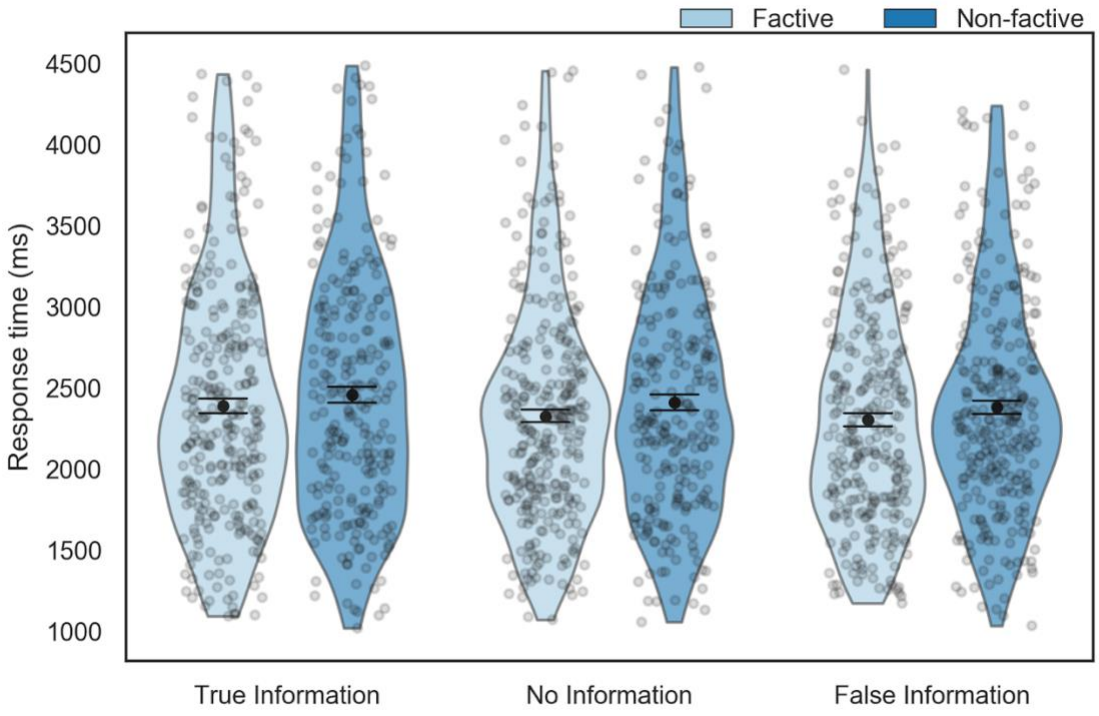


Figure 2.6, Response times for correct evaluations of factive mental state ascriptions (light plots) and non-factive mental state ascriptions (dark plots) as a function of Information Condition. Error bars depict +/- 1 SEM.

2.5.4 Discussion

These results provide evidence that the response time difference observed for evaluations of knowledge and belief may generalize to the broader classes of factive and non-factive mental state verbs to which they respectively belong. This finding supports a more general pattern according to which

factive mental state representations may be simpler because the content represented is necessarily consistent with one's own understanding (Phillips & Norby, 2019). Accordingly, these data also point toward a potential explanation of why evaluations of knowledge occur more quickly than evaluations of belief. However, regardless of whether or not this turns out to be the correct explanation for why evaluations of knowledge are faster than evaluations of belief, the previous four experiments collectively provide ample evidence that people do in fact make evaluations of knowledge without yet having made similar evaluations of belief.

2.6 Experiment 5: Neuroimaging Evidence

In the previous four studies, we investigated the relationship between knowledge and belief by examining the speed with which people were able to evaluate claims about others' mental states. We argued that the fact that people can correctly decide what others know faster than they can correctly decide what others think, provides evidence that they are not consulting their judgments of belief when forming their judgments about knowledge. Here, we ask a similar question but use a very different methodology. Specifically, we use functional magnetic resonance imaging (fMRI) to examine the neural responses exhibited during the formation of knowledge and belief representations and ask whether these neural patterns provide convergent evidence about the relationship between knowledge and belief.

To do this, we took advantage of an existing dataset from an experiment that examined the role of mental state representations in moral judgments for both neurotypical participants and participants with Autism Spectrum Disorder (ASD) (Chakroff et al., 2016). In the original study, participants read short stories involving different kinds of moral violations while undergoing functional magnetic resonance imaging. Critically for our purposes, these short stories manipulated the description of the agent's mental states. In some cases, the agent was described as having knowledge of some fact in the scenario (e.g., "knew the brakes were still broken"); in other cases, the agent was instead described as having a belief

(e.g., “thought the detector just needed new batteries”); and in other cases, the agent was instead described as having some other factive mental state (e.g., “heard that he was in a relationship”; “realized the pond was unsafe”).

While the original research focused on the neural processes underlying moral judgment, we can take advantage of the design used in this study by comparing the patterns of neural activation when participants learned about the agents’ knowledge, belief, and other factive attitudes. Following the literature, we will focus on the neural response within the right temporo-parietal junction (RTPJ), which is well-known to play a highly selective and critical role in the neural computations underlying the representations of others’ mental states ((Gobbini et al., 2007); (Koster-Hale & Saxe, 2013); (Saxe et al., 2004). This area has even been shown, for example, to carry information about the specific kind of mental state being represented; allowing for one to decode whether the agent is represented as having auditory or visual evidence about some fact, among other things (Koster-Hale et al., 2014; Zaitchik et al., 2010).

Our novel reanalysis of these data allows us to test two opposing predictions about the neural activity that knowledge and belief representations will elicit in RTPJ. On the one hand, if representing others’ knowledge requires representing their beliefs, then we would expect the activity within RTPJ to reflect this structure. That is, while knowledge representations may require a number of different or additional neural processes, representations of knowledge should certainly not elicit less of a response in RTPJ than representations of belief, since they are meant to recruit precisely those representations.

On the other hand, the results of the previous five experiments suggest that representations of knowledge may not require representations of belief, and in fact may involve simpler or more basic processes (Phillips, et al., 2020). Previous work has demonstrated that complex or effortful inferences typically elicit greater neural activity in RTPJ than simpler or less effortful inferences ((Cohen et al., 1997); (Meyer et al., 2015)). In line with this, an alternative prediction is that representations of knowledge may elicit a similar amount, or even less, neural activity in RTPJ.

2.6.1 Participants

Participants included 23 neurotypical adults (NT; M_{age} = 27.130; SD_{age} = 11.71; 7 females) and 15 adults with high-functioning autism or Asperger's syndrome (ASD; M_{age} = 31.125; SD_{age} = 8.21; 2 females) recruited from the Greater Boston Area. All ASD participants received their diagnosis based on the Autism Diagnostic Observation Schedule, Second Edition (criterion ≥ 7; M = 9.60), as well as an impression by a trained clinician based on the diagnostic criteria of the DSM-IV.

2.6.2 Stimuli and procedure

As described in Chakroff et al. (Chakroff et al., 2016), participants read 60 stories depicting moral violations while undergoing fMRI. The stories were presented in cumulative segments. The last segment depicted the agent's mental state and was presented for 4 seconds. The agent's mental state was described using three different verb categories: knowledge, belief, or other factive attitude verbs. Word count was matched across conditions. Stories were presented in a pseudorandom order, divided into six 5.5 minute runs. To identify the RTPJ, all participants also completed a theory of mind functional localizer task (Dodell-Feder et al., 2011). This task consists of 10 stories about mental states (false-belief condition) and 10 stories about physical representations (false-photograph condition). (See http://saxelab.mit.edu/superloc.php for the task files). The task was presented in two 4.5 min runs, interleaved with the main experimental runs. Complete details of stimuli and procedure can be found in Chakroff et al. (Chakroff et al., 2016).

2.6.3 Results

As in Dodell-Feder et al. (2011), a whole-brain random effects analysis contrasting neural response in false- belief vs. false-photograph conditions ($p < 0.001$, uncorrected, $k > 16$) revealed activity in the RTPJ (peak voxel MNI coordinates: $x = 58$, $y = -50$, $z = 28$). We identified the RTPJ in all 38 participants individually, defined as contiguous voxels within a 9 mm radius of the peak voxel that passed the contrast threshold.

We averaged the blood oxygen level dependent (BOLD) response across voxels in RTPJ from each subject during the 4 second segment in which the agent's mental state information was presented, and asked if this varied as a function of verb category. For each verb category, we calculated the average percent signal change (PSC) from baseline in RTPJ [PSC = 100*raw BOLD response for (condition – baseline) / raw BOLD response for baseline]. Baseline was defined as the average RTPJ response across all fixation time points between stimuli, adjusted for hemodynamic lag. Complete details of how the fMRI data were processed and analyzed can be found in Chakroff et al. (2016).

We then performed a linear mixed effects analysis of the relationship between verb category (knowledge, belief, other factive attitudes) and RTPJ response using the lme4 package in R (Bates et al., 2015). The full model included random intercepts for subject and item. Including random slopes for the effect of verb category across subjects did not improve the model ($\chi^2(5) = 7.422$, $p = .191$). Furthermore, there was no verb category × group interaction ($\chi^2(3) = 1.209$, $p = .751$), indicating that neurotypical participants and participants with ASD demonstrated similar activity across verb categories.

More importantly, the analysis revealed a main effect of verb category on RTPJ activity ($\chi^2(2) = 9.154$, $p = .010$). Specifically, belief elicited more activity in RTPJ than both knowledge ($t(1784) = 3.043$, $p = .002$) and other factive attitudes ($t(242) = 2.127$, $p = .034$; degrees of freedom calculated using Satterthwaite's approximation) (*Figure 2.7*).

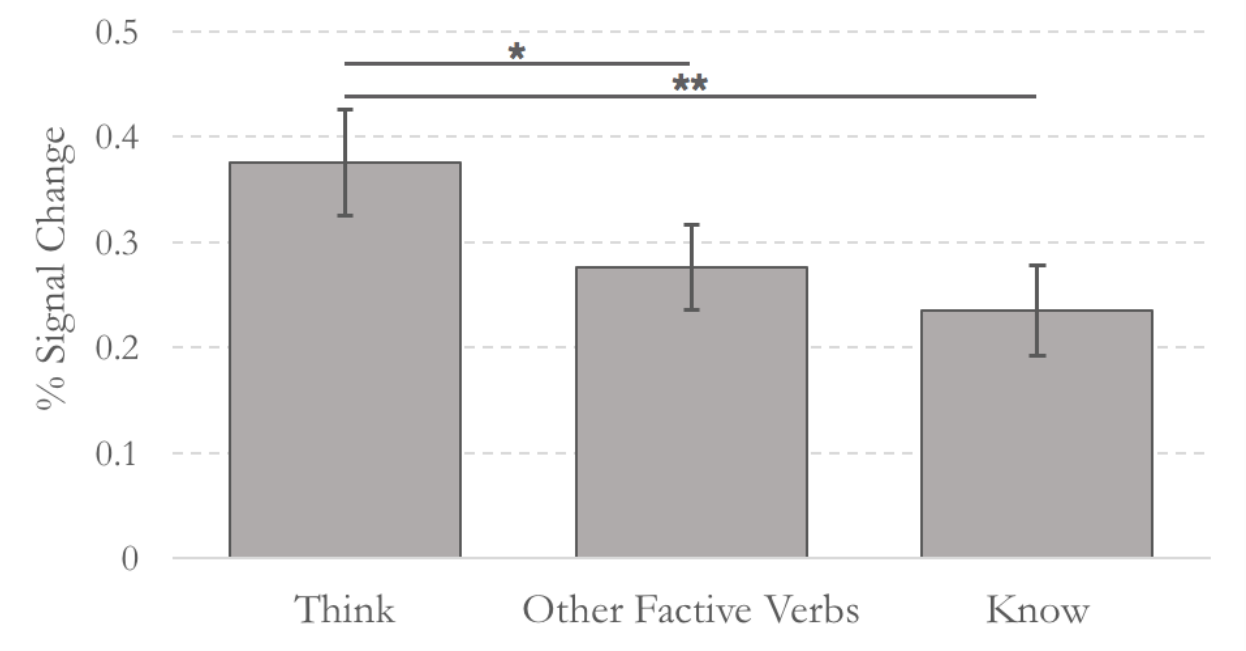


Figure 2.7, Percent signal change in RTPJ for each mental state category: belief (think), knowledge (know) and other factive verbs (e.g., saw, realized). Error bars indicate standard error.

2.6.4 Discussion

Our reanalysis of the fMRI data from Chakroff and colleagues (Chakroff et al., 2016) revealed that RTPJ exhibited a higher level of activity when participants read about others' beliefs than when participants read about others' knowledge (or factive attitudes more generally). This pattern clearly suggests that when participants were representing agents as knowing some proposition, they were not recruiting the set of processes that they employed when representing agents' beliefs. If they had been, then we should have observed at least a roughly similar level of activation in RTPJ. This result is also broadly in line with the findings by Zaitchik and colleagues ((Koster-Hale et al., 2014; Zaitchik et al.,

2010))[7], who found greater activation across the theory of mind network when participants read sentences involving ascriptions of representational mental states (about an agent 'believing' or 'thinking' something) than when they read sentences involving perceptual or emotional states (e.g., that the agent 'saw' something, or what the agent was 'furious' about something). Critically, for our purposes, the perceptual and emotion mental states used were largely factive, while the representational mental states used were largely non-factive ((Koster-Hale et al., 2014; Zaitchik et al., 2010)). Our finding of lower BOLD responses in the RTPJ when representing an agent as knowing as compared to believing is also consistent with the EEG results of Bricker ((2020)), who found reduced inhibition of the P3b amplitude when participants attributed beliefs to another agent as compared to knowledge. Finally, in line with the findings in our Experiment 2, we also do not find differences in the overall pattern observed between neurotypical participants and participants with ASD.

This pattern of results substantially strengthens the case for the conclusions we've drawn across the previous studies. Despite the highly different methodologies used, both approaches yielded surprisingly consistent evidence, which showed that knowledge representations do not rely on belief representations and that knowledge representations may actually be simpler than belief representations. Moreover, the differences between these two approaches complement each other in another way: while alternative proposals may be offered to explain either one of these sets of results, it is very unlikely that any proposal aside from ours can easily explain and predict the pattern of results we observed across both studies (i.e., that representations of knowledge are computed more quickly and require less neural activity in classic theory of mind brain regions).

[7] Zaitchik and colleagues lumped the term 'know' with the representational mental states and did not directly compare knowledge ascriptions to other kinds of representational mental states, as we do here.

2.7 General Discussion

Across five experiments, we found consistent evidence that the representation of others' knowledge occurs in the absence of representations of corresponding beliefs. Experiment 1 demonstrated that accurate evaluations of knowledge ascriptions happen significantly faster than accurate evaluations of otherwise identical belief ascriptions. Experiment 2 further found that this effect cannot be explained by pragmatic differences and that it extends to participants with ASD. Experiment 3 demonstrated this finding cross-linguistically, establishing that people also accurately evaluate others' knowledge faster than their beliefs in French. Experiment 4 provided evidence that this response time difference generalizes to the larger classes of factive and non-factive attitudes to which knowledge and belief respectively belong. Finally, Experiment 5 revealed that forming representations of an agents' knowledge (or other factive mental states) elicits less activity in classic theory of mind brain regions than forming representations of that agent's beliefs. Taken together, these results demonstrate that humans can assess another agent's knowledge without representing their beliefs, and provides support for the more general claim that knowledge representations may be a comparatively more basic form of theory of mind.

2.7.1 Why is knowledge faster than belief?

We take the evidence we've reported to provide support for the claim that the capacity to represent others' knowledge is more basic than the capacity to represent others' beliefs. An important further question concerns what sense in which knowledge is more basic than belief. Given our experimental results, we want to focus on one particularly relevant version of this question, which concerns why people can make accurate assessments of knowledge *faster* than accurate assessments of belief. What is it about knowledge that explains this difference?

One prominent approach in the developmental and comparative literature on mindreading is to propose that knowledge representations can be approximated by some very minimal understanding of others as being connected to particular parts of the world, see, e.g., Butterfill and Apperly's notion of

‘registration’ (Butterfill & Apperly, 2013). However, such ‘minimal’ forms of mindreading are not typically thought to allow for representations of propositional knowledge of the kind tested here, and thus are unlikely to explain why people are faster to correctly evaluate ascriptions of propositional knowledge than ascriptions of propositional belief.

An alternative approach would be to point to the difference in factivity between knowledge and belief. For factive attitudes like knowledge, one can only represent someone as having knowledge when the content of the attitude is consistent with one’s own understanding (Kiparsky & Kiparsky, 2014; Phillips & Norby, 2019). Representations of belief have no such constraint. Accordingly, an intriguing possibility is that this difference in factivity may help to explain why we found that participants could correctly evaluate what others know faster than they could correctly evaluate what others think. One more specific way of spelling out this account, offered by Evan Westra, argues that knowledge representations are *coupled* to the attributor’s own primary representation of reality (Westra, 2021). In Westra and Nagel’s words, “Factive mental state attributions are transparent, in the sense that the attributor looks through them to the world.” (2021, p. 3). In contrast, belief representations are *decoupled* from the attributor’s own understanding, and that maintaining a decoupled representation incurs additional working memory and inhibitory control costs (Fizke et al., 2014), which may explain the slower response times in correct evaluations of belief ascriptions (Westra, 2021). This topic remains an important area for future research.

2.7.2 Interactions between Ascription Type and Information Condition

Across a number of studies (Experiments 1-3), we also found an interaction effect such that the greatest difference in response time in evaluating knowledge and belief ascriptions was in the No Information condition, while a somewhat smaller difference was observed in the True Information

condition, and the smallest difference was consistently observed in the False Information condition. While any proposed explanation of this pattern will necessarily be speculative until further research is done, offering some way of understanding this pattern may still be desirable, given that we observe it relatively consistently.

Thus far, we've argued that correct knowledge evaluations may be faster than correct belief evaluations because evaluations of knowledge may involve simpler or more basic processes. What we now need to explain is why this response time difference is moderated in some cases. One potentially helpful place to start is to recall that knowledge typically entails belief, and so in any case where you recognize someone as knowing something, you should be able to use that fact to infer that they must also believe that thing (Ichikawa & Steup, 2016; Rose & Schaffer, 2013), though see (Myers-Schulz & Schwitzgebel, 2013)). Accordingly, in the True Information condition, where the agent does in fact have knowledge, participants should be able to relatively easily infer that the agent also has the corresponding belief. In contrast, in the No Information condition, when the agent does not have knowledge, participants cannot use this strategy, and they have to independently evaluate whether the agent has the relevant belief. This explanation makes the more precise prediction that the smaller difference in the True Information condition occurs in part because participants are faster to correctly evaluate what the agent thinks, in the True Information condition than in the No Information condition. Our data are consistent with this prediction: combining the data from Experiments 1-3, we found that correct evaluations of belief ascriptions were made faster in the True Information than in the No Information condition, $t(1032) = -3.83$, $p < .001$.

What remains to be explained is why we observe a yet smaller difference in the False Information condition. Importantly, the vignettes used in these conditions explicitly required participants to represent the agent as believing some incorrect proposition, *q*. Recall, however, that we actually asked participants to evaluate whether or not the agent thinks or knows a true proposition *p* (ensuring that the correct answer was the same for both types of ascriptions, as they neither believe, nor do they know, *p*). Thus, the background vignettes used in these conditions require participants to engage in reasoning about what the

agent (falsely) believes, and having previously represented the agent as believing *q*, participants may find it particularly easy to infer that the agent does not believe *p* (as *p* and *q* are always inconsistent). Moreover, determining that the agent does not know *p* may be particularly effortful in this case, as *p* is true, but the agent doesn't know *p* because the agent falsely believes *q* instead of *p*.

While we think this series of explanations is a plausible way of explaining the observed interaction effects, future work should further pursue the potential interaction between the information state of an agent and representations of knowledge and belief.

2.7.3 A difference in kind or magnitude

Our results support the interpretation that the factive mental states of others are relatively easier to represent than non-factive belief states. We proposed that this facility is enabled by the consistency of propositional content between self and other, avoiding the redundancy that may come from decoupling another's view of the world from one's one. However, this explanation may fall short of solving the more complex question of whether the representation of others' knowledge and beliefs constitute categorically different cognitive processes altogether, or if, instead, they result from a single cognitive process carried out to differing extents or magnitudes.

Representations of knowledge and belief exhibit categorically different limitations for the reasons discussed in the general introduction of this chapter. Propositions about the world that are factual, such as historical events, can be both known and believed. While propositions that are false can be believed, but not known. What distinguishes the category of fact from fiction is a philosophical question beyond the scope of this dissertation. Furthermore, there is ongoing debate philosophy concerning whether the state of 'knowing' constitutes a mental state at all (Brueckner, 2002; Williamson, 2002). We take the position here, that knowledge is indeed a mental state whose constituent representations verifiably match the world as it is (Nagel, 2017) .

Butterfill and Apperly have proposed a two-system characterization of mentalizing ability in which an automatic and efficiently fixed system is complimented by a, categorically distinct, deliberative,

and flexible one (2009). If factive and non-factive mental state attributions map onto these distinct systems, then evidence for a difference of kind can come from experimental designs that probe an effect on one system that preserves the second. Additionally, a distinct effect that modulates the second system while preserving the first would, when taken together, provide a double dissociation of factive and non-factive mental state attributions as categorically distinct cognitive processes.

More specifically, identifying a feature of knowledge attribution that is not shared with belief attribution can be used to distinguish the former from the latter. To distinguish the process of belief attribution from that of knowledge attribution would, in turn, require identifying a feature of the former that is not shared by the latter. A candidate feature of knowledge attribution that could distinguish it from belief attribution is automaticity. If task-irrelevant information about another agent's knowledge, but not information about their (even true) beliefs has an interfering or facilitating effect on task performance, for instance on response time measures, this feature would be a sufficient first step to distinguish one process from the other. To achieve the double dissociation needed as evidence for categorically distinct systems, a feature of belief attribution must also be identified that has no effect on the ability to represent others' knowledge. A candidate feature of belief attribution that could distinguish it from knowledge attribution could be the disruptive influence of cognitive load. There is existing evidence that performance on belief reasoning tasks is correlated with tests of general processing speed and executive functioning (German & Hehman, 2006). Suggesting that belief reasoning may be constrained by the speed with which the mind can perform tasks in general. Belief reasoning may also be disrupted when performing simultaneous tasks that interfere with working memory and language (Dungan & Saxe, 2012; McKinnon & Moscovitch, 2007; Newton & de Villiers, 2007). It remains to be seen whether stresses on these more deliberative and effortful moderators of belief reasoning have a similar influence on participants' ability to reason about others' knowledge states. Identifying the features of a double dissociation between mentalizing processes involving knowledge and belief states is an important topic for future research.

2.7.4 Conclusion

This paper set out to ask whether people can represent someone else as knowing something without representing them as believing that thing. Across a series of five studies using a variety of manipulations and methods, we found strong evidence for an affirmative answer to this question: representations of what others do or don't know cans occur in the absence of representations of what they believe. This series of findings provides further support for the growing consensus that the capacity for knowledge representation is more basic than the capacity for belief representation.

Chapter 3.

3.1 Introduction

An open question in research on causal cognition concerns the relationship between causal reasoning about ordinary physical events and causal reasoning about agents. For example, do we use the same cognitive processes when deciding that a falling tree caused damage to the car as we do when deciding that a CEO's poor decisions caused damage to the company stock? On one side, there is a long history of work focused on how humans reason about outcomes brought about by physical objects (Michotte, 1946). A strength of these designs comes from the computational tractability of the laws constraining the operation of physical objects. Humans possess internal models capable of simulating the mechanics of rigid bodies in space. When considering what physical objects caused an outcome, this intuition constrains the list of culpable candidates to only those objects that *could* be causal in accordance with the laws of physics. For instance, we are unlikely to consider objects that do not make physical contact with a balloon as possible causes of the balloon bursting since that is not how the physics of balloon bursting events typically operate. In humans, infants as young as six months old demonstrate clear evidence of an appreciation for concepts in physics (Liu & Spelke, 2017; Liu et al., 2017; Ullman et al., 2017). The intuition for physical causality among objects has even been documented at the sensory level, where past work has revealed that photoreceptors of the retina show an adaptation effect to causal events that makes subsequent ambiguous events more likely to be seen as non-causal (Kominsky & Scholl, 2020; Rolfs et al., 2013). In machines, the laws of physics can be represented abstractly through mathematical formulations programmed into simulation engines and applied in a range of contexts from civil engineering to virtual reality gaming. This mechanistic understanding of events has aided cognitive scientists studying causal reasoning in purely physical contexts and has begun to offer compelling

accounts of causal judgments in these cases. However, an important question is how such causal reasoning about relatively simple physical events is related to causal reasoning about more complicated events that involve animate agents who choose which actions to perform.

Another, relatively separate, approach to studying causal cognition has focused on these more complicated cases and includes variables more like those we observe in real-life situations. These designs often include verbal descriptions of scenarios involving humans intentionally bringing about outcomes in various ways and ask participants for explicit judgments of causality, often in complex moral contexts (Alicke, 1992; Driver, 2008; Samland & Waldmann, 2016). A strength of these approaches comes from their ecological validity. In this sense, the insights they yield have direct applicability to the problems in the real world that hinge on the ability to assign responsibility. However, the ecological validity enjoyed by vignette approaches to studying how humans attribute causality to others may come at the expense of construct validity. As such, methods to study causal attribution to agents remain difficult to characterize 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).

Although researchers have typically used distinct methodologies to study the cognitive processes recruited when reasoning about physical objects and intentional agents, there is little compelling evidence to suggest these processes rely on distinct cognitive mechanisms. The approach taken in the following studies attempts to bridge these separate literatures to contribute towards 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 physical laws.

In Experiment 1, we manipulate the animacy of a candidate cause as either goal-directed or objectively inanimate. By holding the fine-grain kinematics of the actual and counterfactual outcomes fixed across animate and inanimate conditions, we can isolate the effect of perceived animacy on causal attribution judgments. One important feature distinguishing agents from objects is referred to as *equifinality*, wherein the same outcome can obtain, despite variation in the means. In contrast, inanimate

objects exhibit *multifinality* such that a variation in the means produces variation in the outcome. Both of these properties point to a difference in the counterfactual outcomes that are possible for agents and objects. Despite this potentially important difference in causal reasoning about agents and objects, we find nearly identical patterns of causal judgments about both entities in the first experiment. In Experiment 2 we sought to explore whether causal judgments of the two entities may come apart when we vary the relevance of counterfactuals involving agents and objects. We collected causal and counterfactual judgments for a physically overdetermined outcome, in which either of two candidate causes, alone, would be sufficient to cause the outcome. We test if any differences in causality assigned between agents and objects for an overdetermined outcome might be mediated by the counterfactuals we consider in each case. While the brain can detect animacy virtually automatically (Schultz et al., 2005), what is often less certain are the underlying intentions that cause agentic behavior. Consistent evidence has accrued to show that descriptive and prescriptive norm violations have a strong influence on causal judgments (Hitchcock & Knobe, 2009; Icard et al., 2017; Morris et al., 2019). Furthermore, deliberate human actions tend to be judged as more causal of outcomes than unintentional human mistakes (Fincham & Jaspers, 1980; Lagnado & Channon, 2008; McClure et al., 2007), leading to challenges in disentangling judgments of causality from judgments of moral responsibility (Kominsky & Phillips, 2019; Sytsma et al., 2012). In Experiment 3 we parse these variables apart by looking at the role played by intentional or unintentional prescriptive norm violations on causal judgments of anthropomorphized agents. Taken together, we hope that this work narrows the divide between the way causal attribution to humans and non-human causes is commonly conceptualized.

3.2 Experiment 1: The influence of animacy on causal attribution.

3.2.1 Background

This experiment investigates whether intuitive knowledge of causal relationships generalizes from objects to animate agents. Prior work provides some reason to think that causal cognition may operate differently for animate agents vs inanimate objects. For example, people who cause outcomes deliberately are seen as more causal than people who cause outcomes unintentionally (Hilton et al., 2016; Hilton & Slugoski, 1986; Lombrozo, 2010; Malle et al., 2014; Samland & Waldmann, 2016). Accordingly, since inanimate objects always lack intentionality in their movement, it is possible that they may be considered less causal than a goal-directed agent for bringing about the same outcome.

Alternatively, the claim that causal judgments for goal-directed agents and inanimate objects operate under the same cognitive mechanisms is supported by accounts of counterfactual relevance (Gerstenberg et al., 2021; Kominsky & Phillips, 2019) and broader dependence theories of causality (David Lewis, 1973, 1974). According to counterfactual accounts, attributions of causality to some candidate, either a physical object or mental agent, crucially rely on evaluations of whether or not the outcome would obtain in counterfactuals in which the candidate was altered or removed (Danks, 2017; Lipe, 1991).

In the current study, we investigate whether the "behavior" of animate, goal-directed agents are judged as causes to the same extent as the movement of inanimate objects for the same outcome. Crucially, we isolate the influence of perceived animacy on causal judgments by holding the physical parameters of the causal events and outcome fixed. If causal cognition operates differently for goal-directed agents vs inanimate objects, we expect to see differences in causal attribution judgments between animacy conditions within participants. Alternatively, if the cognitive mechanisms underlying these judgments are instead isomorphic, animated agents and inanimate objects will be judged as equally causally responsible for the same outcomes. Furthermore, we investigate whether the relationship between counterfactuals and causal judgments will hold in the same way with respect to agents and objects.

3.2.2 Methods

Participants

105 adults were recruited from Prolific (app.prolific.co) to participate in Experiment 1 (71 males). All participants were at least 18 years old, (M_{age} = 24.99, SD_{age} = 8.227), endorsed fluency in English, and had a ratio of successful task submission above 94%. All participants were required to complete the task on a personal desktop computer or laptop.

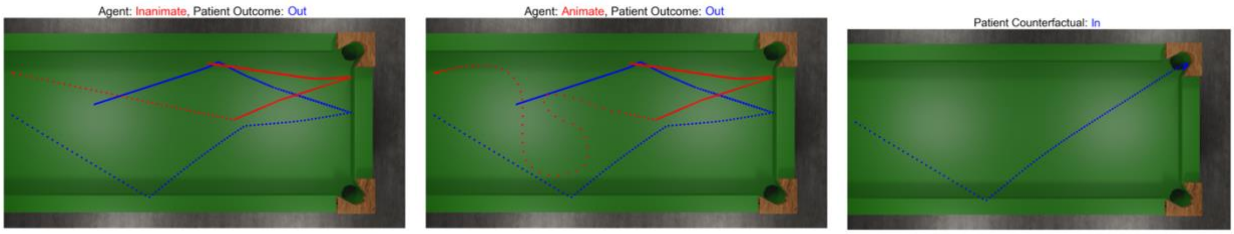
Stimuli and Procedure

Before discussing our investigation further, it is important to note how we use specific terminology, as this may be a source of confusion for readers. The classic empirical research on causal cognition focused on the perception of "launching" events in which one object appears to collide with another static object, thereby "causing" its subsequent motion (Michotte, 1946). In these cases, and in the experiments to follow the object that launches is referred to as the *agent*, whereas the object that *is* launched is referred to as the *patient.* In what follows, we use these terms accordingly, i.e. 'The agent exerts a force on the patient'. Where appropriate, however, we will refer to agents as being either animate or inanimate.

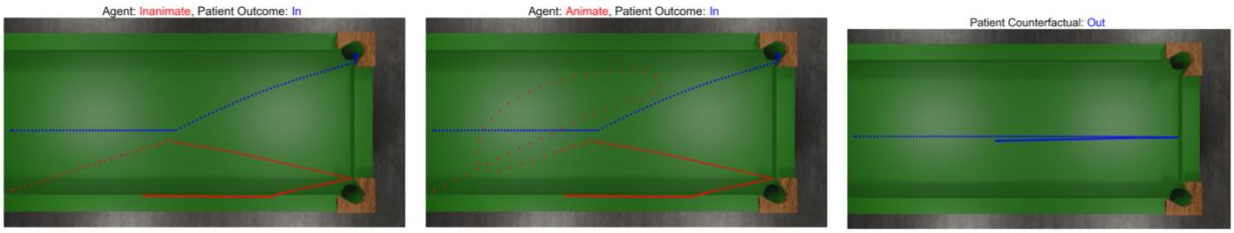
Stimuli depicted a billiard ball table in which a red ball ("Ball A") collides with a blue patient ball ("Ball B") resulting in the patient landing in or missing one of two corner pockets. Along with the animacy of the agent, we manipulated whether the patient ball would have landed in or missed the corner pocket in the counterfactual where the agent ball is removed from the scene. 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. We created six unique patient trajectories for each combination of patient outcome, and patient counterfactual outcome (qualitatively varying the degree to which the patient's counterfactual outcome differs from its actual outcome) resulting in 24 unique patient trajectories. Each patient trajectory was tested with an animate and inanimate agent for a total of 48 video

clips presented to each subject. Four (two animate and two inanimate) of these video clips were included as catch trials, in which the agent does not make contact with the patient. All stimulus videos in this and subsequent experiments were created using Blender 3D computer graphics software v2.9, which uses the Bullet physics engine to simulate ball trajectories and collisions. Crucially, the trajectory and fine-grain kinematics of the patient ball were cached by the physics engine so that they could be saved and matched across animacy conditions. In videos containing animate agents, agent trajectories were manually specified along a bezier curve roughly simulating the movement of an animate, goal-directed agent and culminating at the location, angle, and with an instantaneous velocity equivalent to the complimentary inanimate agent for the same patient trajectories (*Figure 3.1*, A-D).

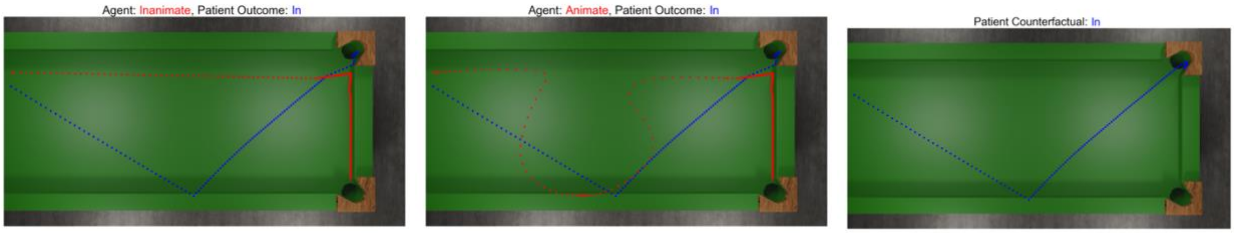
A.



B.



C.



D.

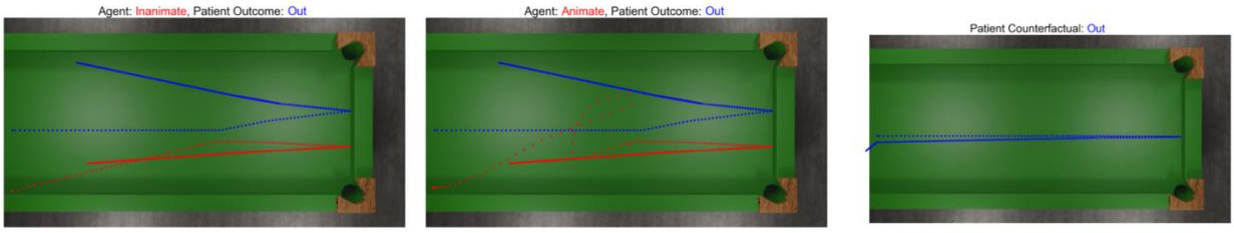


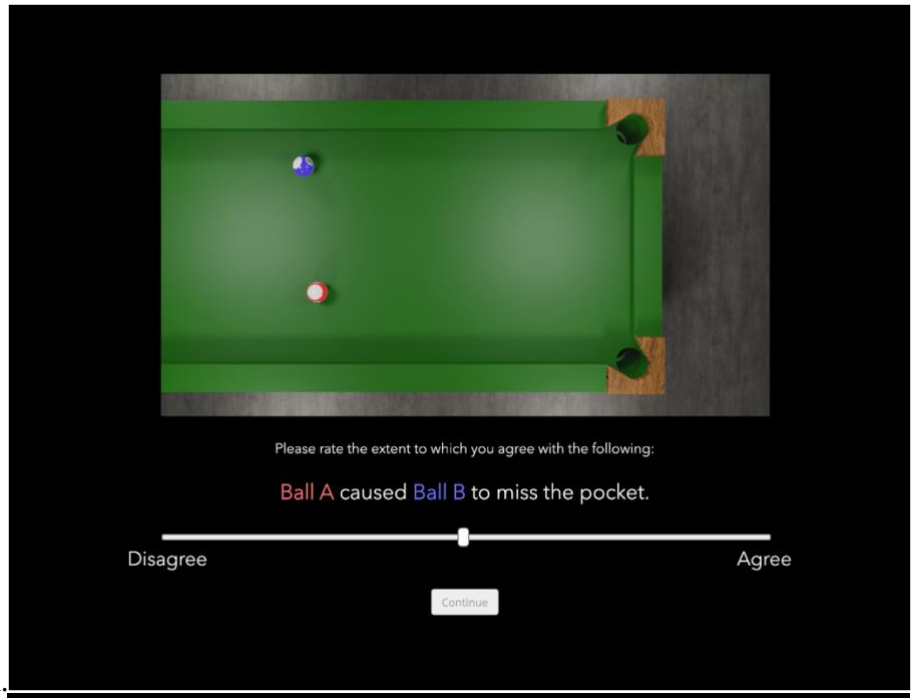
Figure 3.1, Examples of stimulus conditions. Scatterplots depict the xy-location of the agent (red) and patient (blue) ball in each frame of the given video clips (closer dots indicate slower movement). The left and center columns of each row depict scenarios with inanimate and animate agents respectively. The right column of each row depicts the relevant patient counterfactual physics simulation in which the agent ball is removed (not viewed by participants). Rows represent each combination of actual and counterfactual patient outcomes of landing in or out of a corner pocket. ***A****. Agent makes a difference in patient outcome, causing the patient to land out, which would have counterfactually landed in.* ***B****. Agent makes a difference in patient outcome, causing the patient to land in, which would have counterfactually landed out.* ***C****. Agent collision does not make a difference in patient*

outcome, the patient landed in after collision, but would have counterfactually landed in without agent intervention. ***D****. Agent collision does not make a difference in patient outcome, the patient landed out after collision, but would have counterfactually landed out without agent intervention.*

The stimulus used in this, and all other experiments reported here, was delivered through a custom web application built using the Flask python microframework and hosted on custom-built encrypted servers (www.linode.com). We modified plugins from the jsPsych toolbox to suit the needs of the current study (de Leeuw, 2015). After consent, participants were instructed to use slider bars on the screen to make judgments about a series of billiard ball events based on a short video clip.

In each trial, an eight-second video clip was presented depicting a billiard ball table in which a red ball ("Ball A") collides with a blue patient ball ("Ball B") resulting in the patient landing in or missing one of two corner pockets. The video controls were removed and the video played on a loop for the duration of the trial. After the clip was played at least once in full, a slider bar appeared and participants were asked to rate the extent to which they agreed with the prompt shown below the video. Depending on the outcome of the patient, the prompt asked participants to indicate their agreement with the statement "Ball A caused Ball B to land in [miss] the pocket". Ratings were made on an integer scale from 0 ("Disagree") to 100 ("Agree")(*Figure 3.2A*). Note that the agent, which collides with the patient can be animate or inanimate. Across all trials, the patient ball ("Ball B") was always inanimate. After indicating their agreement with the causal statement, participants indicated the extent to which they perceived the agent ball ("Ball A") as animate (*Figure 3.2B*).

A.



B.

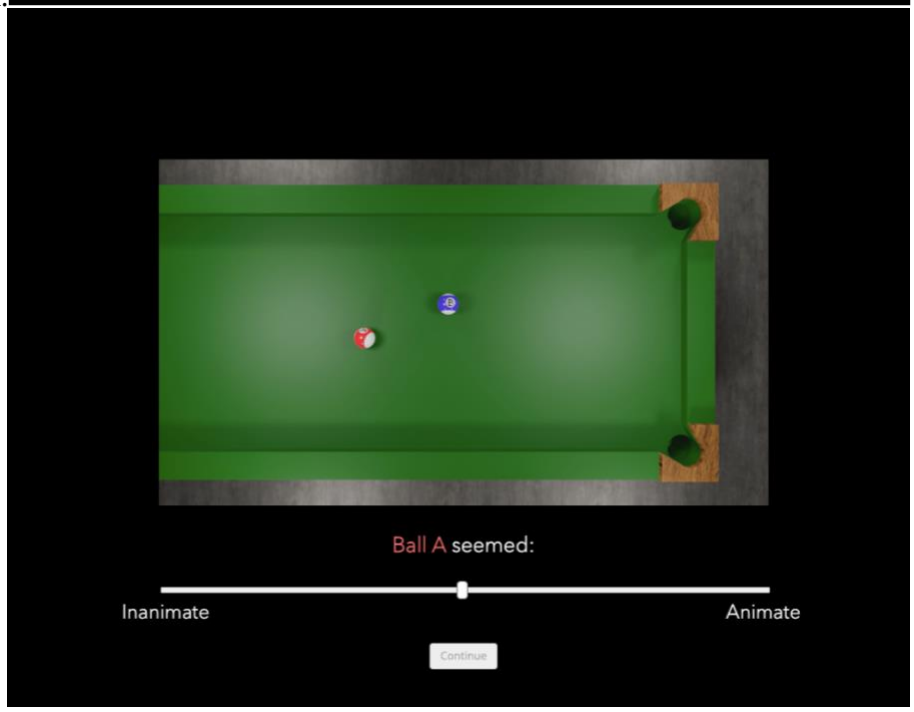


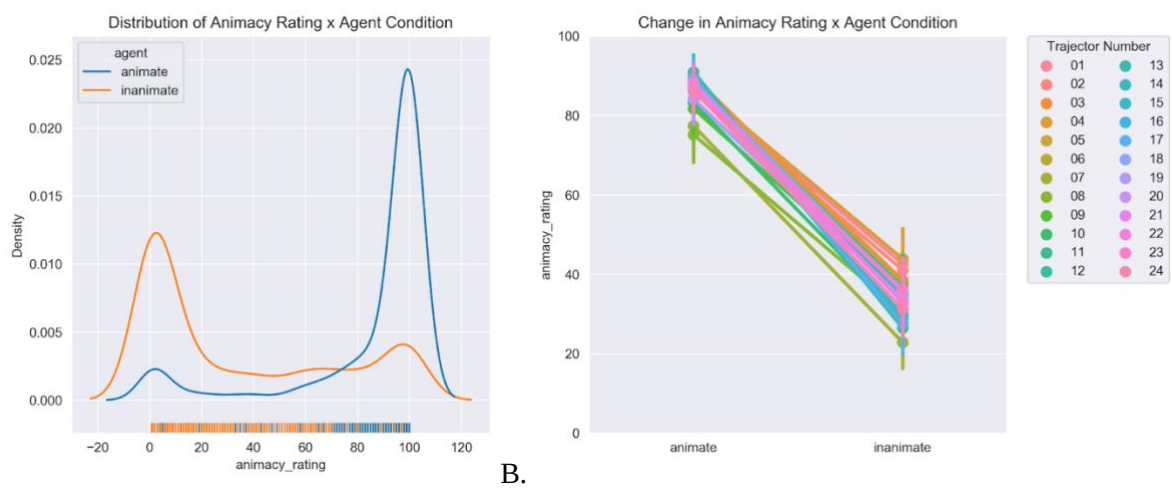
*Figure 3.2, Example trial **A**. Causal statement for the agent ball endorsed by participants. **B**. Animacy judgment of the agent ball.*

All participants observed and made causal and animacy judgments for all 48 stimulus clips, before being debriefed on the nature of the study and compensated.

3.2.3 Results

Participants were excluded from analyses if they endorsed an agreement rating of ≥ 60 with the causal statement in catch trials in which the agent and patient do not make contact. For all remaining participants, these trials were excluded from further analysis.

Before conducting the main analysis, we found that our manipulation of animacy was indeed successful across participants. The results of a likelihood ratio test reveal a large difference in animacy rating between trials in which agents' movements were manually manipulated to appear goal-directed ($M_{animate}$ = 86.27, $SD_{animate}$ = 27.85), and trials in which the agent's movement was rendered from a physics simulation ($M_{inanimate}$ = 33.80, $SD_{inanimate}$ = 37.85), $\chi^2(1) = 103.76$, $p < .0001$ (*Figure 3.3A-B*).



A. B.

*Figure 3.3, Animacy manipulation, **A**. Distribution of judgment ratings for agent ball's animacy by animacy condition. **B**. Change in mean animacy rating for the agent within each unique patient trajectory (i.e. difference in animacy judgment between agent balls represented in the left and center columns with each row of Figure 3.2).*

First, we tested for a three-way interaction effect of patient outcome x patient counterfactual x agent animacy on participants' causal attribution rating for the agent using random effects for subject and patient trajectory. This did not reveal a significant three-way interaction effect on causal ratings, $\chi^2(1) = .2539$, $p = .6144$. Next, we tested for the two-way interaction effect of patient outcome and patient counterfactual outcome on causal ratings for the agent across animacy conditions. Consistent with existing theories (Woodward, 2003), we found the interaction of patient outcome and patient counterfactual significantly influence causal ratings for the agent, $\chi^2(2) = 11.322$, $p = .0007$. Planned pairwise comparisons of the six possible outcome/counterfactual combinations were carried out using the Estimated Marginal Means package in R (Lenth et al., 2022). These tests reveal that the interaction we observed was driven by causal judgments on trials in which a patient's outcome is changed from counterfactually in to actually out by an agent across both animacy conditions. More specifically, participants rated both animate and inanimate agents as more causal when it diverts the patient from landing in the corner pocket, changing the patient outcome from counterfactually in to out ($M_{in \rightarrow out} = 80.73$, $SD_{in \rightarrow out} = 31.33$), than when the patient outcome is unaffected by the agent and lands out ($M_{out \rightarrow out} = 44.46$, $SD_{out \rightarrow out} = 43.86$), $t(27.28) = 2.94$, $p < .05$, or lands in ($M_{in \rightarrow in} = 45.64$, $SD_{in \rightarrow in} = 41.2$), although this difference only approached significance, $t(26.81) = -2.43$, $p = .09$ (*Figure 3.4*). In other words, participants found agents that make a difference in the patient's outcome more causal, regardless of whether the agent was animate or inanimate. This effect was greater when patients ultimately missed the corner pocket, than when patients ultimately land in.

Finally, we tested for the main effect of agent animacy on causal ratings. We found no effect of agent animacy on causal ratings across all outcome and counterfactual conditions $\chi2(1) = 0.0228$, $p = .88$, (Figure 3.4, blue vs. orange).

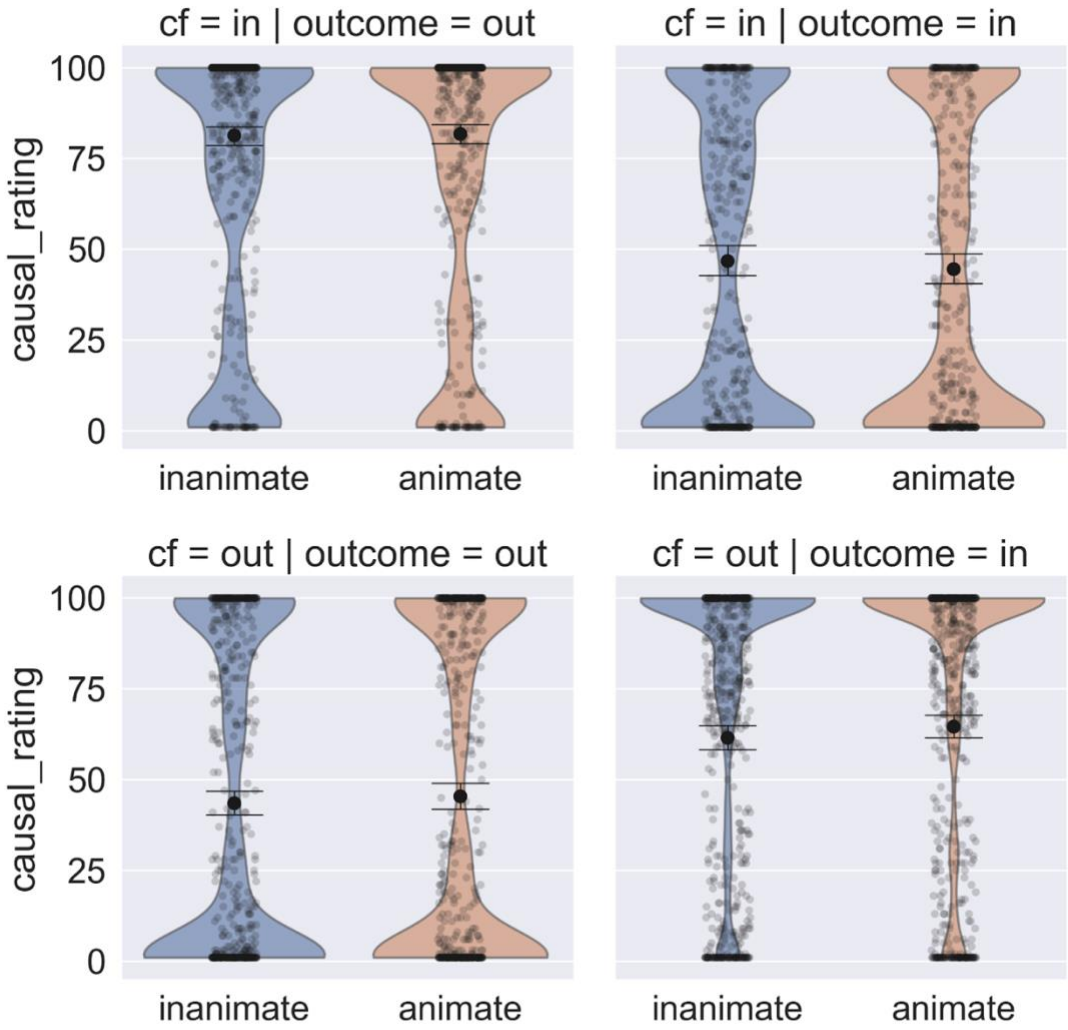


Figure 3.4, Causal ratings for inanimate and animate agents across counterfactual and outcome combinations. Error bars depict ± 1 SEM.

3.2.4 Discussion

The purpose of this experiment was to test whether goal-directed agents are judged as more responsible than inanimate objects for causing the same outcomes. Although there is evidence to suggest that people use distinct forms of causal reasoning for deliberate and unintentional causal factors, we

reasoned, instead, that similar processes could be employed with respect to inanimate objects and goal-directed agents when holding other variables fixed. We created stimuli to test whether causal cognition for intentional agents and inanimate objects operate under unified or distinct mechanisms. In a classic collision context, we manipulated the animacy of the agent, as well as the patient's actual and counterfactual outcomes, and asked participants the extent to which they believed the agent (either animate or inanimate) caused the patient's outcome. Importantly, everything about the fine-grain kinematics of the outcomes and counterfactuals were matched, such that the only difference between animacy conditions was the agent's movements *prior* to the frame in which the collision occurred. We found that, despite clear differences in the perceived animacy of the cause, participants made roughly equal attributions of causality. As expected, we found that the interaction of the patient's actual and counterfactual outcome had a significant effect on causal ratings for the agent ball. Furthermore, this interaction was most clearly driven by the condition in which the patient's outcome was changed from counterfactually landing in the pocket to missing it post-collision. This result was somewhat surprising. Because there are far fewer ways of hitting the patient into the corner pocket than diverting it away from the corner pocket, we expected the highest causal ratings to occur for animate or inanimate agents that changed the patient outcome from a miss to landing in. This would capture the fact that the agent was more clearly a difference maker when the ball went in rather than missed. Instead, we found the opposite: the agent was judged as more causal when it caused a ball to miss that would have counterfactually gone in. A potential reason for finding this pattern could be an artifact of the billiard game context we used. Typically, the goal of the game is to hit balls into the table pockets. Therefore, an agent that changes a patient's outcome from a positive one in the context of the game to a negative one, or unexpected one may influence causal judgments. Prior work has demonstrated that value and expectations have strong effects on causal cognition (Icard et al., 2017; Knobe & Fraser, 2008).

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 is consistent with previous studies demonstrating that causal cognition for

intentional agents and inanimate artifacts may be underpinned by the same mechanistic accounts (Kominsky & Phillips, 2019). We extend this finding here and provide an even more compelling case that includes objects animated to appear intentional. In the current experiment, patient counterfactuals were held constant to explore the effect of agency in isolation. We show that animacy, itself, does not impact causal judgments directly. In Experiment 2, we consider if outcomes are viewed as inevitable (i.e. lacking alternative counterfactuals) if caused by inanimate objects, or intentional agents.

3.3 Experiment 2: The influence of animacy on causal attribution via counterfactuals

3.3.1 Background

In Experiment 1, we show that animacy, in isolation, does not make a candidate more causal of an outcome according to adult participants. One possible reason we did not observe a difference in causal judgment is that patient counterfactuals were matched across animate and inanimate conditions. In Experiment 2, we indirectly manipulate the contrast of actual and counterfactual outcomes through the perception of animacy. Keeping outcomes matched across conditions, we examine if animacy influences whether or not an outcome is viewed as inevitable. We expect that viewing an agent's intention vs. an object's physical trajectory as the cause of an outcome will change the possible counterfactuals that may be considered such that the outcome *could* be different given a different intention. We focus on the case of overdetermination, in which an outcome's occurrence depends on any of multiple, individually sufficient causal events. A conceptually similar paradigm was used by Walsh and Sloman, who presented participants with vignette scenarios in which two agents simultaneously throw rocks at a bottle. The authors establish a causal structure known as late preemption, in which one rock is described as reaching the bottle before the other. In these cases, participants will judge it more causal to the outcome (Walsh &

Sloman, 2011). However, when the scenario is described such that both rocks hit the bottle simultaneously, causal strength is shared equally (Chang, 2009). These overdetermined outcomes are interesting case studies because they can disentangle *processes* theories of causality, which require that some quantity is physically transferred from cause to effect, from *dependence* theories, which require different outcomes in counterfactual contrasts (Paul, 1998). Both rocks transfer some quantity to the bottle, however, the bottle breaking does not depend on either rock individually. Gerstenberg et al.'s Counterfactual Simulation Model provides compelling evidence that causal attribution judgments involve online simulations of counterfactual possibilities (2021). For overdetermined outcomes of physical causes, it is easy to imagine, in accordance with process accounts of causation, two causes having equal strength when occurring simultaneously and the earlier cause judged as more causal in the case of late preemption. Here, we investigate cases in which the sufficient causes of an outcome include an inanimate object or a goal-directed agent. When an outcome occurs via the intentions of an agent or the force of an object, only the agent holds the *potential* to change the result. Thus, we expect that any effect of perceived animacy on causal judgment will be mediated by the difference in counterfactual outcomes considered for the animated agents and inanimate object agents. As such, in a disjunctive structure where an animate or inanimate agent ball is individually sufficient for the outcome, the inanimate agent will be judged less causal for the same outcome than an animate one, who is capable of changing the outcome if they intended. The mechanism for this relies on the counterfactuals that come to mind such that 1.) The outcome will be viewed as overdetermined in the case of two objects; defined by the lack of possible counterfactuals in which either candidate can make a difference in the outcome; 2.) The outcome will not be viewed as necessarily overdetermined in the case when an agent ball is perceived as animate. In the latter scenario, the agent's intentions play a role in the final outcome, and therefore a relevant counterfactual exists in which the intention is removed, and the animate agent moves in order to change the outcome.

In this experiment, the outcome event judged by participants is held fixed across animacy conditions once again. It depicts a scenario in which two causes bring about an overdetermined outcome.

However, the perceived animacy is manipulated via a brief priming clip between participants, thus changing the counterfactual possibilities.

3.3.2 Methods

Participants

210 adults were recruited through Prolific for participation (M_{age} = 38.23, SD_{age} = 14.40, 107 females). 106 participants were randomly assigned to the animate agent group, the other 104 participants were randomly assigned to the inanimate agent group. All participants endorsed fluency in English and had a ratio of successful task submissions above 94%.

Stimuli and Procedure

As in Experiment 1, the stimulus was delivered through a custom web application hosted on an encrypted remote server. After consenting, participants were instructed that they would be watching two short video clips of different events and that we would be asking them to make judgments afterward.

To manipulate the perception of animacy, a priming clip was displayed before the test clip. For participants randomly assigned to the inanimate agent condition, the priming clip depicted two balls rolling into view on a platform that also contained an assembled tower of cubes. The balls were shown colliding with each other and bouncing off different edges of the platform (*Figure 3.5A*). Movements of both balls in the inanimate agent condition and one 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 a static patient ball and an animated agent ball that appeared to be "playing" with the inanimate patient, repeatedly knocking it around the platform, chasing it, and colliding with it again, avoiding the assembled tower of cubes also present on the platform (*Figure 3.5B*). Movement of the agent ball in the animate condition was manually specified along a bezier curve approximately simulating 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 an assembled tower of cubes. This time, the inanimate patient ball rolls into view headed straight for the tower of cubes. 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 all crashing down. Importantly, both balls in the test clip were rendered according to a physics simulation, such that the test clips were identical across the animate and inanimate conditions. (*Figure 3.5C*). The video controls were removed and the video played on a loop while participants made causal and counterfactual judgments.

Participants were asked to indicate their agreement using a slider scale ranging from 0 ("totally disagree") to 100 ("totally agree") with two causal and two counterfactual statements, counterbalanced across participants. For causal items, participants endorsed the statement "*The [agent color / patient color] ball caused the tower to fall.*" For counterfactual items, participants endorsed the statement "*If the [agent color / patient color] ball had not been there, the tower would have remained standing.*" After making their ratings, all subjects completed a short comprehension check where they indicated whether or not one of the balls appeared animate, as well as which of the two balls made first contact with the tower before being debriefed.

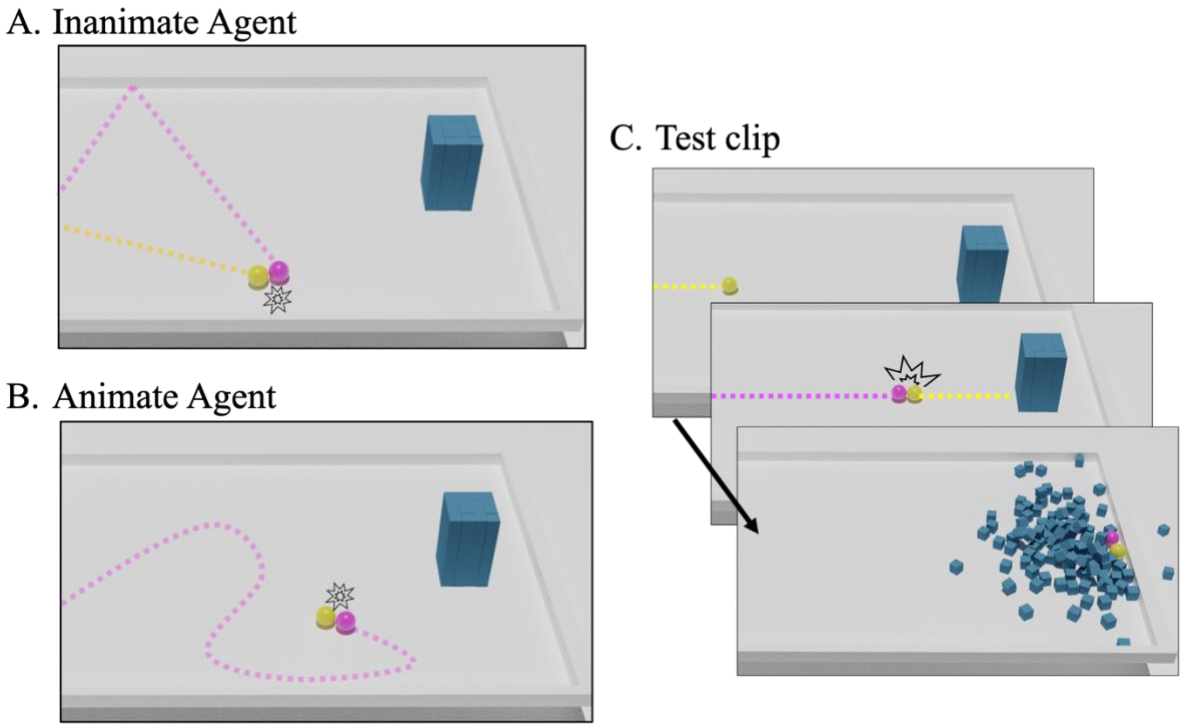
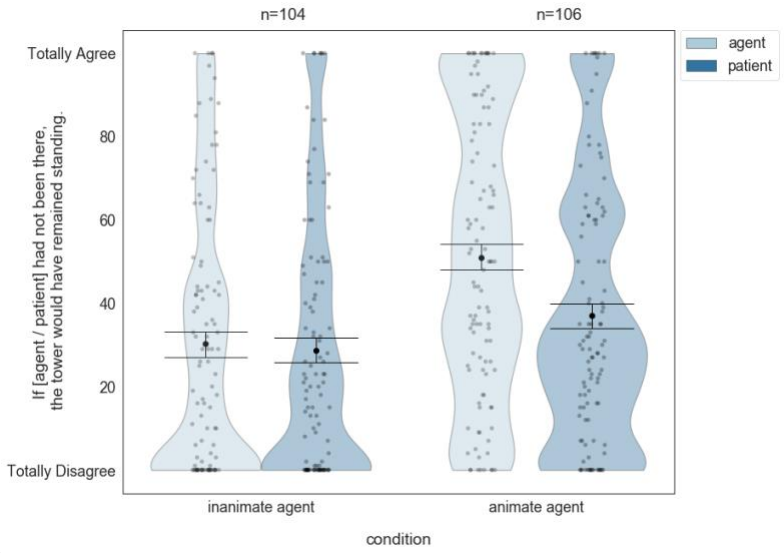


Figure 3.5, Frames of the video stimuli used in experiment 2. Star shapes indicate a collision between balls. Dotted lines illustrate the trajectory of each ball. ***A****. The priming clip shown to participants in the Inanimate agent group. Movements of both balls were rendered from a physics simulation engine.* ***B****. The priming clip shown to participants in the Animate agent group. Movements of the pink ball were manually specified along a bezier curve roughly simulating the movement of an animate agent "playing" with a ball. The movement of the yellow ball was rendered from the physics simulation engine and was responsive to the force generated by the animate agent.* ***C****. The test clip viewed by participants in both agent groups. First, the yellow patient ball rolls into view headed straight for the tower of cubes. Shortly after, the pink agent ball rolls into the frame along the same trajectory and collides with the yellow patient ball, which proceeds to crash into the tower of cubes, bringing it all crashing down. Movements of the pink agent, yellow patient ball, and cube tower were rendered from the physics simulation engine. Participants were asked to endorse causal and counterfactual claims about each ball for the tower collapsing.*

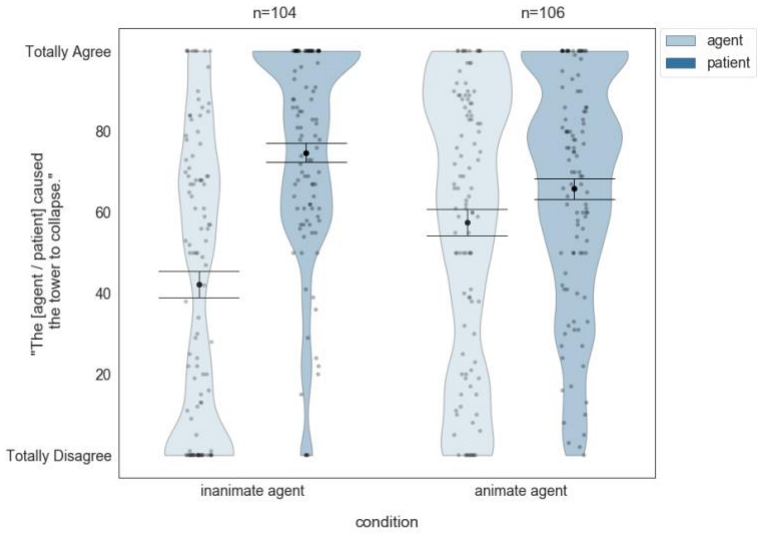
3.3.3 Results

As we anticipated, there was a strong effect of counterfactual dependence judgments on causal rating, when controlling for animacy $F(1, 208) = 72.55$, $p < .0001$. However, participants in the Inanimate Agent group provided low counterfactual dependence ratings for both agent ($M_{\text{inanimate}} = 30.38$, $SD_{\text{inanimate}} = 31.42$) and patient ($M_{\text{inanimate}} = 28.71$, $SD_{\text{inanimate}} = 30.83$) balls. Given the criteria for causality according to process and dependence theories, this result is consistent with the predictions described above for cases involving overdetermined outcomes. Interestingly, we find the highest counterfactual dependence ratings for the agent ball in the Animate Agent group (*Figure 3.6A*). This influence of agent animacy on counterfactual dependence ratings was highly significant such that the outcome was judged more counterfactually dependent on animate agents ($M_{\text{animate}} = 51.00$, $SD_{\text{animate}} = 33.19$) than inanimate agents ($M_{\text{inanimate}} = 30.38$, $SD_{\text{inanimate}} = 31.42$), despite both causes following identical trajectories in the test clip, $F(1, 208) = 21.39$, $p < .0001$ (*Figure 3.6A*, light plots).

We hypothesized that there would be a difference in participants' causal judgments of animate vs. inanimate agent balls for the same outcome. Using a one-way analysis of variance to test for the total effect of animacy on agent causal ratings, we found that perceived animate agents indeed were rated significantly more causal ($M_{\text{animate}} = 57.61$, $SD_{\text{animate}} = 33.54$) than inanimate agents ($M_{\text{inanimate}} = 42.19$, $SD_{\text{inanimate}} = 35.27$) for the same outcome event, $F(1, 208) = 10.54$, $p = .00136$ (*Figure 3.6B*, light plots). Interestingly, causal ratings were highest for the *patient* ball in the Inanimate Agent condition ($M_{\text{patient}} = 74.88$, $SD_{\text{patient}} = 26.45$), suggesting that something other than counterfactual dependence may be driving causal judgments when both balls are inanimate and the outcome is overdetermined.



A.



B

Figure 3.6, (A)Counterfactual and (B) causal ratings for agent and patient balls in the Animate Agent and Inanimate Agent between-subjects conditions. Error bars depict ± 1 SEM.

Using the mediation package in R, we found that the indirect (average causal mediation) effect of animacy on causal ratings was (β_{animacy} = 20.64)*($\beta_{\text{counterfactual}}$ = .542) = 11.19. We tested the significance of this indirect effect using nonparametric bootstrapping procedures. The 95% confidence interval of the bootstrapped unstandardized indirect effect ranged from 6.077 to 16.58. Thus, the indirect effect of animacy via counterfactual dependence judgments on the agent ball's causal ratings was statistically significant, $p < .001$ (*Figure 3.7*). In contrast, this indirect effect did not explain causal judgments made for the patient ball, 95% CI: [- .816, 2.07], $p = .560$.

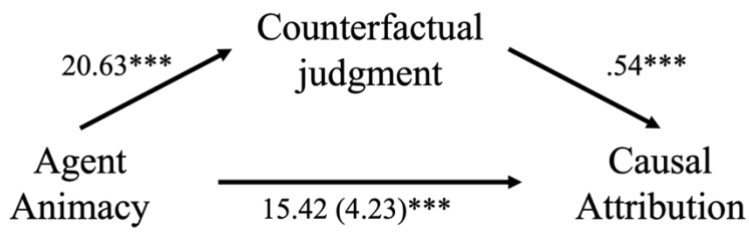


Figure 3.7, Mediation relationship of the direct and indirect effect of animacy on causal attribution ratings for the agent ball.

3.3.4 Discussion

Here, we examined the prediction that an effect of perceived animacy on causal judgments could be explained by the mediating influence of counterfactuals. We reasoned that the perception of animate agency makes counterfactuals in which an agent makes a difference in the outcome (i.e. by having a different intention) relevant for consideration in the causal attribution process. By contrast, we expected that an outcome caused by either of two inanimate objects in the overdetermined scenario we tested would be interpreted as inevitable by participants such that there are no counterfactuals in which the outcome could be different. To test this, we presented two participant groups with the same test clip in which a patient ball rolls into view headed straight for the tower of cubes. 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 all crashing Participants in the Animate Agent group were primed with an anthropomorphic stimulus, while those in the Inanimate Agent group were primed using a purely physical stimulus. We asked participants to make causal judgments for each candidate ball. We also asked participants the extent to which the outcome counterfactually depended on either ball.

We found that, when the agent ball was viewed as inanimate, judgments of counterfactual dependence for both causes were low. This result makes sense given that the outcome under consideration was overdetermined and therefore 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, despite participants viewing identical outcomes in the test clip across groups. Although the test clip was rendered from a physics simulation, this could suggest that participants did not interpret the outcome as inevitable when the agent ball was perceived as animate. Furthermore, the pattern was mirrored in participants' causal attributions. Distal agent balls were judged as more causal to the outcome when perceived as animate rather than inanimate.

We found the expected pattern of results in which counterfactual judgments of difference-making were strongly related to causal attribution judgments across animacy groups. Finally, the effect of perceived agent animacy on causal strength judgments was mediated by counterfactual judgments of whether the outcome would obtain in the absence of the agent.

In study 3, we look closer at the role of mental states and the perception of animacy in considerations of causal judgments by exploring the role of prescriptive norms on counterfactuals.

3.4 Experiment 3: Dissociating animacy from prescriptive norms in causal attribution.

3.4.1 Background

In Experiment 1, we show that goal-directed animacy alone does not impact causal judgments. Next, in Experiment 2, we found that animacy *does* impact causal judgments, but only to the extent that it arbitrates which counterfactual possibilities might be considered. Here, we examine how the relationship between prescriptive norms and counterfactual considerations influences causal judgment. Counterfactuals are challenging to investigate in animate agents because, along with their observable actions, an agent's goals, desires, and beliefs may also be considered causal to events and are free to be virtually anything. Using norm violations to study causal cognition partially overcomes the combinatorial explosion of possible counterfactuals and the variables therein ("the variable selection problem" (Beebee, 2004; Bernstein, 2015; Hesslow, 1988)) by making salient the counterfactual in which the norm is instead followed (Petrocelli et al., 2011). Here, we manipulated both animacy and normativity to examine how prescriptive expectations are at the core of determining counterfactual relevance for animate agents.

Consistent evidence has demonstrated that people have a tendency to attribute increased causality to agents that violate social or moral norms as compared to agents who do not (Henne et al., 2019; Kirfel & Lagnado, 2018; Knobe, 2009; Kominsky et al., 2015). However, differences in judgments of causal strength have also been reported between agents acting deliberately and those who unintentionally cause a negative outcome (Hilton et al., 2016; Hilton & Slugoski, 1986; Lombrozo, 2010; Malle et al., 2014; Samland & Waldmann, 2016).

In Experiment 2, animacy was used to cue participants to the possibility of alternative outcomes given a counterfactual intention of an animated agent. We found that perceived animacy affects causal judgments indirectly through the counterfactuals relevant to goal-directed agents. In the current

experiment, we take advantage of prescriptive norm violations to manipulate precisely which counterfactuals participants are likely to consider. More specifically, we manipulate the perception of animacy in participants with a priming clip again. However, this time we vary participants' normative expectations of the animate agent such that, in bringing about a destructive outcome, the agent is viewed as violating one of two kinds of prescriptive norms, moral or rational. We use a similar outcome event to that of Experiment 2, only now, the agent and patient are both perceived as animate by some groups of participants. We predict that anthropomorphized agents will be considered more causal to an outcome when violating participants' normative expectations.

3.4.2 Methods

Participants

All participants were recruited through Prolific. The eligibility criteria for this experiment matched those of Experiments 1 & 2, with the additional filter to exclude Prolific users who had participated in either of the above experiments. 587 Participants were recruited in two cohorts. 293 adults (M_{age} = 32.65, SD_{age} = 12.48, 148 females) were recruited to participate in the causal judgment cohort, and 294 adults (M_{age} = 32.41, SD_{age} = 11.93, 146 females) were recruited to the counterfactual judgment cohort. Both cohorts were split among the same stimulus conditions and differed only in the judgment data we elicited.

Stimuli and Procedure

As in Experiments 1 and 2, the stimulus for this experiment was delivered through a custom web application hosted on an encrypted remote server. After consenting, participants were instructed to watch two short video clips of different events and were asked to make judgments afterward. Participants assigned to either of two animate conditions viewed a priming clip involving two animated agents. In the animate priming clip, participants observed a platform containing two different colored balls, one pink and one green ("the builder") as well as several cubes partially assembled into a tower, and individual cubes at various locations on the platform. Both balls were animated to appear self-guided and goal

directed. In the video, the green "builder" ball is shown moving to different cubes and assembling the tower one cube at a time, while the pink ball moved interactively as if observing the green ball build the tower (*Figure 3.8A* & *C*). Importantly, a description providing context for the priming clip was varied by condition. Participants in the Animate Immoral condition viewed a message above the priming clip informing them that "It is Pink's job to protect Green's tower" (*Figure 3.8A*). Participants in the Animate Irrational condition viewed a message above the priming clip informing them that "Green wants to protect its own tower" (*Figure 3.8C*).

Participants assigned to the inanimate condition viewed a different priming clip containing an inanimate green ball and an inanimate pink ball for consistency with the animate conditions. Each ball in the inanimate priming clip rolled into view on a platform that also contained an assembled tower of cubes. The balls were shown colliding with each other and bouncing off different edges of the platform (*Figure 3.8E*). Movements of both balls in the inanimate agent condition were simulated using the Bullet physics simulation engine. Participants in the Inanimate condition were simply asked to familiarize themselves with the video, as they would be asked about a similar one on the subsequent screen.

After the priming clip, participants in all conditions viewed a similar test clip. The test clip was similar to the test clip of Experiment 2 in terms of the underlying physics. It contained the same platform and an assembled tower of cubes. It depicted the patient ball rolling into view headed straight for the tower of cubes. Shortly after, another 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 all crashing down. Importantly, both balls in the test clip were rendered according to a physics simulation, such that the ball movements and the tower outcome in the test clips were identical across conditions. (*Figure 3.8B, D, & F*). The video controls were removed, and the video played on a loop while participants made causal or counterfactual judgments. Critically, we manipulated the color of the agent and patient balls in the test clip between conditions in order to violate the norm established in the prime clip. Participants in the Animate Immoral condition watched the pink "guard" collide with the green "builder" patient, violating the moral prescription established in the prime clip for the pink "guard" to protect the tower (*Figure*

3.8B). Participants in the Animate Irrational condition observed the green "builder" collide with the pink patient ball, violating the rational prescription established in the prime clip of the green "builder" wanting to protect its own tower (*Figure 3.8D*). In the inanimate condition, the color of either ball was set randomly (*Figure 3.8F*).

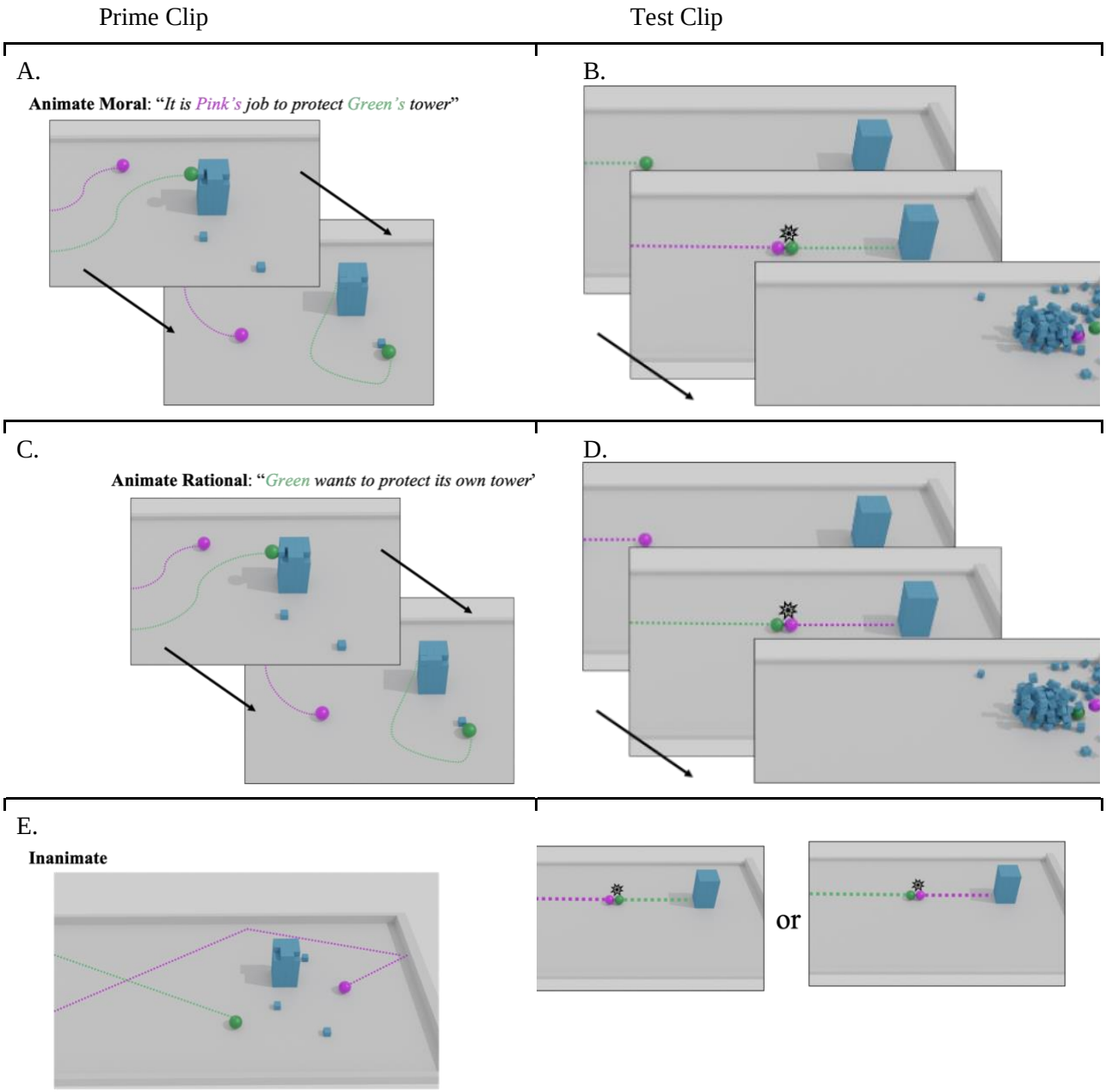


Figure 3.8, Frames from the prime and test stimulus videos. A. Animate Immoral condition: The green agent assembles the tower while the pink agent observes interactively. A prescriptive is established from the text. B. The moral prescription is violated by the pink agent in the test clip. C. Animate Irrational condition: the prime clip is the same as in A, but a rational prescriptive is established. D. The rational prescriptive is violated by the green agent in the test clip. E. Inanimate condition: Two inanimate balls roll into view colliding with each other and bouncing off edges, not making contact with the tower. F. The test clip in this condition matches the animate condition with the color of the agent and patient balls chosen randomly.

Causal & counterfactual judgments

Participants in the first cohort made causal attribution judgments to both balls for the tower's outcome by rating the extent to which they agreed with the statements "[Agent / Patient] caused the tower to fall" using a slider with values ranging from 0 ("Disagree") to 100 ("Agree"). Participants in the second cohort made counterfactual judgments for both balls for the tower's outcome by rating the extent to which they agreed with the statement "If [Agent / Patient] had not been there, the tower would have remained standing" using a slider with values ranging from 0 ("Disagree") to 100 ("Agree"). Additionally, participants in the counterfactual cohort also indicated their agreement with the statements "I expected [Agent] to move in a different way than it did in the video.", and "If [Agent] had moved in a different way, the tower would have remained standing.".

3.4.3 Results

Causal judgments

We began by analyzing the effect of each condition on causal judgments of the agent ball. We found that, 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 3.9*, light plots). Pairwise comparisons carried out using the Estimated Marginal Means package in R (Lenth et al., 2022) revealed that the agent ball in the Animate Moral condition was judged as more causal ($M_{\text{immoral}} = 66.21$, $SD_{\text{immoral}} = 29.31$) than the agent ball in the Animate Irrational condition ($M_{\text{irrational}} = 41.57$, $SD_{\text{irrational}} = 32.78$) for the same outcome, $t(289) = 5.36$, $p < .0001$. The agent in the Animate Moral condition was also judged more causal to the outcome than the agent in the Inanimate condition ($M_{\text{inanimate}} = 40.41$, $SD_{\text{inanimate}} = 34.13$), $t(289) = 5.61$, $p < .0001$. Interestingly, causal judgments of the agent ball in the Animate Irrational condition were not significantly different from the Inanimate condition, $t(289) = 0.250$, $p = .97$.

Recall that the tower outcome was physically overdetermined in all conditions such that the tower collapsing did not counterfactually depend on either agent or patient in the physics simulation. Nonetheless, causal attribution to agents and patients differed from each other in each condition. As

expected, the patient, which actually makes first contact with the tower, was judged more of a cause ($M_{patient}$ = 72.97, $SD_{patient}$ = 25.23) than the agent ($M_{inanimate}$ = 40.41, $SD_{inanimate}$ = 34.13), in the Inanimate condition $t(96) = 6.03$, $p < .0001$, paired (*Figure 3.9*, right). Surprisingly, this was also true for the patient ($M_{patient}$ = 71.65, $SD_{patient}$ = 28.48) in the Animate Irrational condition, $t(96) = 5.25$, $p < .0001$, paired (*Figure 3.9*, center). As we hypothesized, this pattern reverses in the Animate Immoral condition, where the agent ($M_{immoral}$ = 66.21, $SD_{immoral}$ = 29.31) is judged more causal than the patient ($M_{patient}$ = 55.96, $SD_{patient}$ = 32.22) (*Figure 3.9*, left). However, this difference only approached significance $t(96) = -1.83$, $p = .06$, paired.

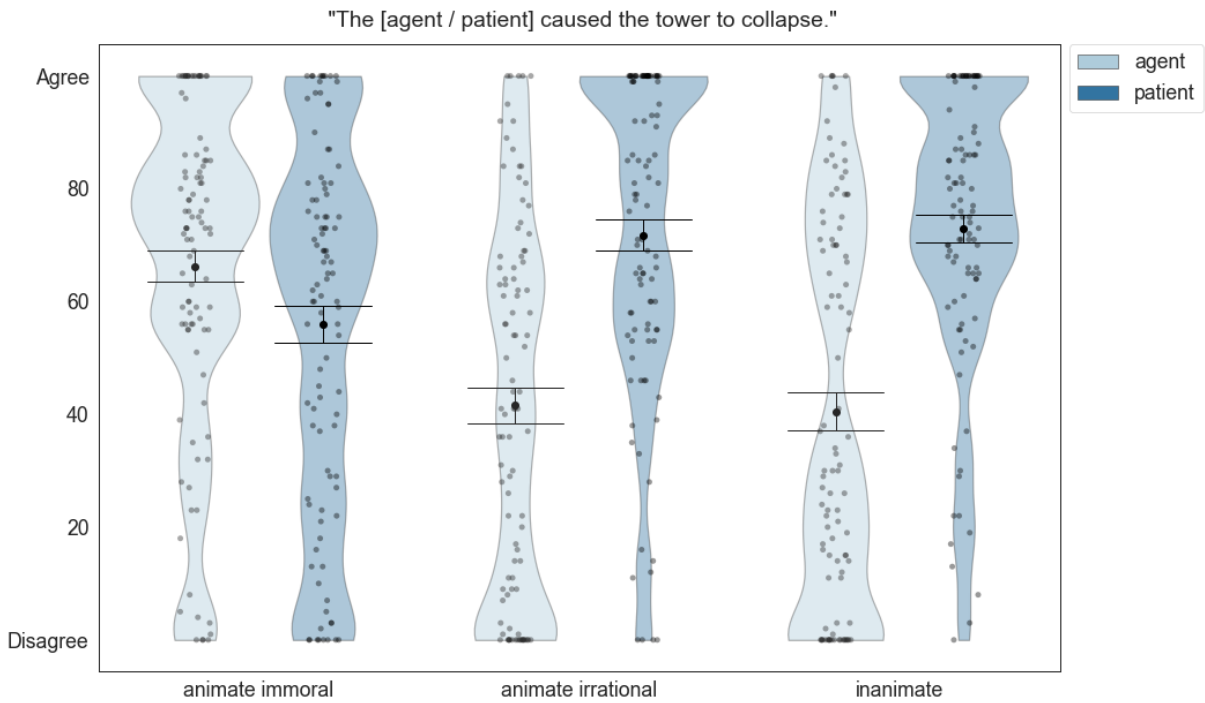


Figure 3.9, Causal attribution judgments for agent (light) and patient(dark) balls across conditions.

Counterfactual judgments

We found a significant interaction effect of condition x ball (agent /patient) on counterfactual dependence judgments $\chi^2(2) = 54.31$, $p < .0001$. This interaction was driven by the Animate Irrational condition, in which participants judged the tower collapsing as more dependent on the patient ($M_{patient}$ = 69.37, $SD_{patient}$ = 34.82) than the agent (M_{agent} = 34.03, SD_{agent} = 34.37), $t(297) = -7.56$, $p < .0001$ (*Figure 3.10A*, middle). 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 Animate 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 ($M_{immoral}$ = 78.32, $SD_{immoral}$ = 26.72; $M_{irrational}$ = 71.22, $SD_{irrational}$ = 28.86) were found more surprising than inanimate ones($M_{inanimate}$ = 30.19, $SD_{inanimate}$ = 29.42), $F(1, 292) = 160.31$, $p < .0001$. There was no difference in the extent to which agent "behavior" violated participant expectations between immoral and irrational agents $t(291) = 1.75$, $p = .190$ (*Figure 3.10B*).

Finally, to test if participants believed the outcome counterfactually depended on the agent's surprising movement, we asked the extent to which they agreed that the tower's outcome would be different if the agent 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, $SD_{immoral}$ = 29.27; $M_{irrational}$ = 63.07, $SD_{irrational}$ = 28.75) moved differently than if the inanimate agent had ($M_{inanimate}$ = 35.37, $SD_{inanimate}$ = 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 3.10C*).

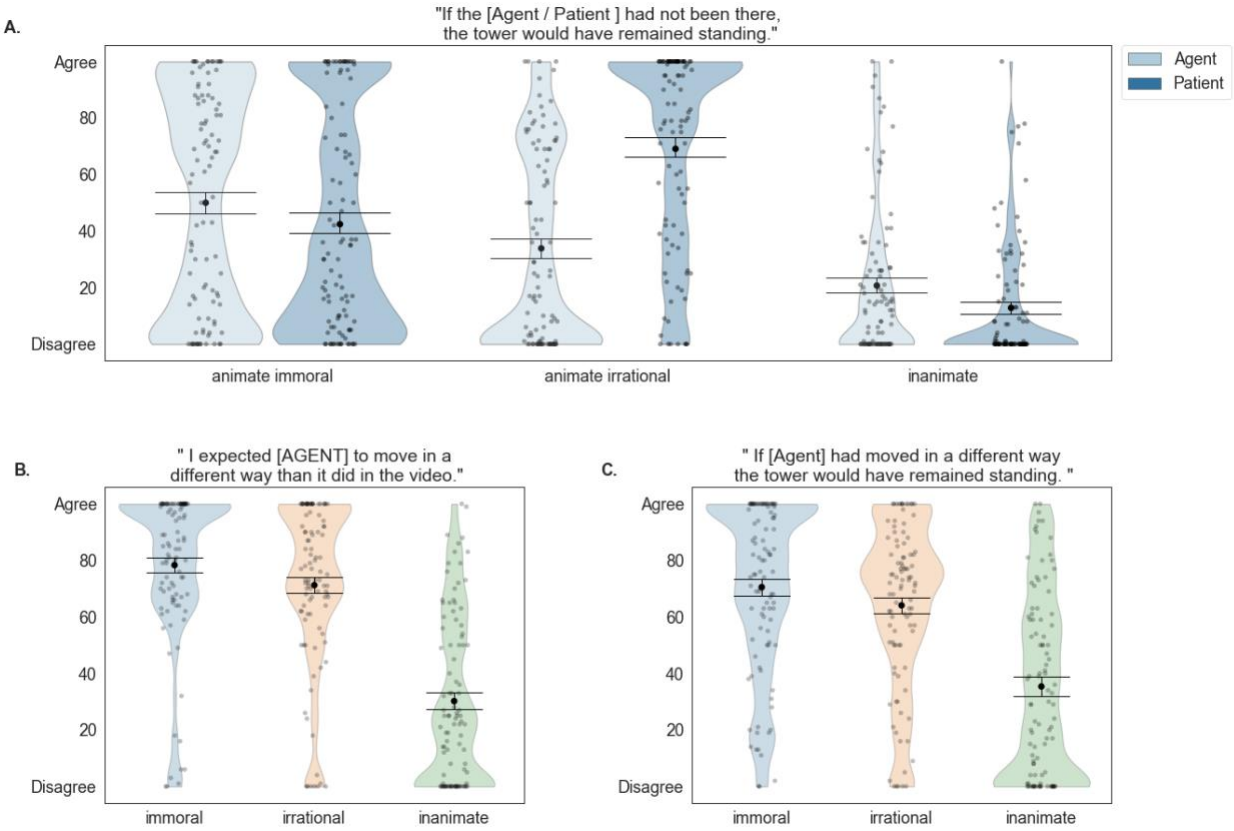


Figure 3.10, A. Participants' judgments of the extent to which the outcome counterfactually depended on the agent (light) and patient (dark) balls respectively. B. The extent to which the movement of the agent violated participant expectations. C. Participants' beliefs that the outcome would be different if the agent had moved in a different way.

3.4.4 Discussion

Norm violations are known to influence causal judgments in a variety of interesting ways (Hitchcock & Knobe, 2009; Icard et al., 2017). One reason norm violations provide such rich testbeds for researchers is because of the known importance of counterfactuals for causal cognition (David Lewis, 1973). In the context of humans who cause outcomes, prescriptive norm violations have often been operationalized in terms of morality (Alicke et al., 2011; Leslie et al., 2006; Samland & Waldmann, 2016). By contrast, prescriptive norm violations when studied with respect to causal objects are often

described as norms of proper functioning (Hitchcock & Knobe, 2009; Kominsky & Phillips, 2019; Lombrozo, 2010). In this experiment, we were interested in the effect of prescriptive norm violation on causal judgments of anthropomorphized agents. We manipulated the perceived animacy as well the normativity of a causal agent and found that animated "immoral objects" were judged as more causal to a destructive outcome than their inanimate counterparts.

We also asked whether or not moral norms are special cases for causal cognition by also presenting subjects with a rational norm violation in which an agent behaved inconsistently with their expressed interests. In both immoral and irrational conditions, we expected the animated agents to be judged as more causal of the outcome than their inanimate counterpart given the results in Experiment 2, where this was the case for a similar outcome event. Interestingly, causal attribution ratings for these irrational agents were significantly *lower* than those of immoral agents (*Figure 3.9*). To understand this surprising result, we turned to the counterfactual judgements made by the second participant cohort. For both irrational and immoral between-subjects conditions, participants agreed that the agent's movements violated a norm, and that the outcome depended on this norm violation. However, there were larger differences between immoral and irrational conditions in agreement with the counterfactual statement about whether the outcome would obtain in the absence of the agent or patient. More specifically, there were unexpectedly higher agreement ratings that the outcome was more counterfactually dependent on the presence of the *patient* than the "irrational" agent (*Figure 3.10A*, middle). In other words, participants did not believe that our irrational "builder" agent, alone, would have harmed the tower it had just completed. We speculate that this likely came from a fortuitous error of errors, so to speak. It is likely that, due to the aspects of our stimuli, participants did not perceive the animated "builder" as acting *with the intention of* making the tower collapse. Although the outcome in the test clip was confirmed as overdetermined in the physics simulation engine such that the agent alone would, in fact, have caused the tower to fall, these data suggest that participants viewed our "irrational" agent as simply causing the outcome by mistake--perhaps even in the process of trying to prevent the other ball from hitting the tower.

Our unconvincing stimulus was fortuitous in the sense that it provided a more interesting dissociation between deliberate and unintentional prescriptive norm violations for agentic objects.

A common criticism of work on causal attribution in contexts where prescriptive norms are violated is that participants may conflate questions of causality with questions of moral responsibility or blameworthiness. Phillips and Kominsky point out that when actions that violate prescriptive norms are made unintentionally, for example, by agents who are ignorant of the prescriptive norm, they may not be considered moral violations since there is no normative expectation that an agent would abide by prescriptions if, e.g. they lack critical knowledge of them (2019). In the current experiment, participants likely viewed our "irrational" agent as causing the tower's destruction accidentally, roughly embedding a condition of intentionality within our condition of norm violations by animate agents.

We take up the recent definition of intentionality offered by Quillien and German:

> "For the human mind, an agent did X intentionality if the agent's attitude toward X caused X, and caused X according to the typical causal model implicit in our commonsense psychology" (2021).

An agent's attitude in this definition refers to the value they place on some outcome, X, obtaining and includes the agent's tolerance for the costs associated with bringing about X. Quillien and German underscore that simply having a desire for X is insufficient for the claim that an agent causes X intentionally if X comes about through a means that does not comport with our commonsense understanding of the general way in which X is caused (e.g. Despite having the *desire* to score a point, a basketball player who closes their eyes and takes a shot would not be judged as *intentionally* scoring a point if their shot lands in the basket by coincidence). Consistent with this idea, Sosa et al provide compelling evidence for a distinction between moral and causal reasoning. Participants in their study viewed a display in which an anthropomorphized agent exerts varying amounts of effort to cause harmful outcomes for others. The authors found that a model including parameters for the agent's *physical* causal contribution to the outcome did not improve predictions of participants' *moral* judgments above a simpler account that modeled the morality of the agent as a function of its perceived *desire* to cause harm (Sosa et

al., 2021). Other recent work has shown that the epistemic state of an agent matters more than the abnormality of their actions when considering an agent's causal contribution to an outcome. For example, if two agents must both perform an action in order to jointly cause a single outcome, whether or not one agent knows of the other's behavior make a difference in causal attribution judgments (Kirfel & Lagnado, 2021).

Overall, the general pattern of causal attribution results fits well with the pattern of results found for counterfactual dependence judgments in a separate cohort. Taken together, this experiment provides support to the claim that the underlying mechanism by which prescriptive violations, both deliberate and unintentional, affect causal cognition likely stems from a more domain-general effect of normativity.

3.5 General Discussion

In Experiment 1, we asked whether the property of being 'alive' and causing an outcome 'intentionally' was sufficient, in itself, to impact participants' causal judgments. We found that, when holding both realized and counterfactual outcomes fixed, goal-directed animacy, alone, had no effect on causal judgments. Instead, we replicate prior work demonstrating the causal strength attributed to agents and objects alike is a function of counterfactual difference-making (Danks, 2017; Woodward, 2003). In Experiment 2 we reasoned that outcomes need not be considered inevitable if an agent possesses the ability to intervene in a way that objects cannot. This allows the counterfactual possibilities for agents to differ from those of objects. We show that this difference in counterfactuals mediated the differences in causal strength attributed to agents and objects for an outcome with the same underlying physics. Finally, we examined the relationship between perceived animacy and counterfactual relevance more closely in Experiment 3. Although we cannot reverse-engineer the behaviors of goal-directed agents to compute which mental states are hypothetically possible, observing norm violations highlights those counterfactuals which are more probable. This gives us the most relevant counterfactual contrasts for free.

We found higher causal attributions to perceived “immoral objects” who violated prescriptive norms. Incidentally, we replicated prior work demonstrating that causal attributions are higher for deliberate norm violations than unintentional ones. This is because we failed to elicit the perception of irrationality in participants, who instead, may have viewed an animate agent’s actions as unintentional and therefore less causal. While the perception of animacy allows for the possibility of different outcomes given different intentions, our data suggest that the counterfactual contrast participants considered did not involve an alternative intention (to protect its tower in accordance with rationality), but instead involved an alternative behavioral process by which the rational intention could be carried out successfully. What this suggests is that the influence of intentionality on causal attributions to agents may be better explained by an observer's normative expectations.

A gap has emerged in the research on causal cognition between the questions concerning our understanding of causal objects and agents. This rift parallels that between philosophical schools of thought concerning how to define causation. For decades, collision events between inanimate agents have been used to bolster processes theories, which promote the notion that causation is strictly defined by the physical transference of some quantity between cause and effect (Aronson, 1971; P. Dowe, 2000; Salmon, 1994; Wolff, 2007). On the other side, some have argued that what separates cause from correlation is the effects of interventions on causal, but not correlative relationships (Pearl, 2000; Woodward, 2003). As such, human agency to cause outcomes places an emphasis on our mental states in ways described by dependence theories, that promote a requirement of counterfactual dependence between cause and effect (e.g. letters appearing on this page are caused by my desire to type them here; since without the intervention of my desires, these letters would not be here). We believe distinct accounts of causal cognition for agents and objects would be just as superfluous as the rift between process and dependence theories of causality more broadly.

Exciting new work is bridging this philosophical divide by incorporating aspects of process and dependence to characterize causal judgments for purely physical events (Gerstenberg et al., 2021) with compelling evidence from human psychophysics (Gerstenberg et al., 2017). The success of Gerstenberg et

al.'s counterfactual simulation model stems, in part, from the fact that human intuition of what is physically possible in realized or counterfactual worlds is constrained by a programmable set of laws. The mental states that cause our behaviors are far less understood; and much less so, programmable. In this set of experiments, we sought to unify various accounts of causal cognition by imbuing objects with animacy. By comparing anthropomorphized agents to objects in this way, we could control the physical contributions to an outcome, while observing the effects of perceived mental agency on causal attribution.

Chapter 4.

4.1 Introduction: Causal Selection

An understanding of causal relationships gives humans the amazing power to intervene on their environment; changing it to suit their needs. Knowledge of causal relationships is learned from the experience of a continuous stream of temporally ordered events and allows us to explain the past and predict the future. It also plays a pervasive role in our society, for instance in legal or medical contexts. However, investigation of the mechanisms involved in causal reasoning has mostly been studied using coarse manipulations. Existing work in experimental philosophy using simple vignette-based experiments has only just begun to scratch the surface of this complex topic. Norms, both statistical and prescriptive, have consistently been shown to influence how humans attribute causality. Evidence for the effect of normality has come mostly from work manipulating this variable in broad, qualitative ways by describing events as either normal or abnormal. (Halpern & Hitchcock, 2013; Hitchcock & Knobe, 2009; Icard et al., 2017). However, we believe normality can and should be construed in less binary ways. In fact, recent work has already demonstrated that the effects described above persist along more continuous, quantitative manipulations of normality in simple two-variable conjunctive and disjunctive structures (Morris et al., 2019). The purpose of the causal selection benchmark dataset is to provide researchers with a comprehensive set of human causal judgment data to test the predictions of various computational and theoretical models that propose to explain the influence of normality on how humans reason about the causes of events.

4.1.1 Causal selection: What we are studying, and what we are not

Knowledge of causal relata can be represented in various forms. *Token* causality forms links between specific events that occur only once (*c* caused *e*). On the other hand, *type* causation refers to non-

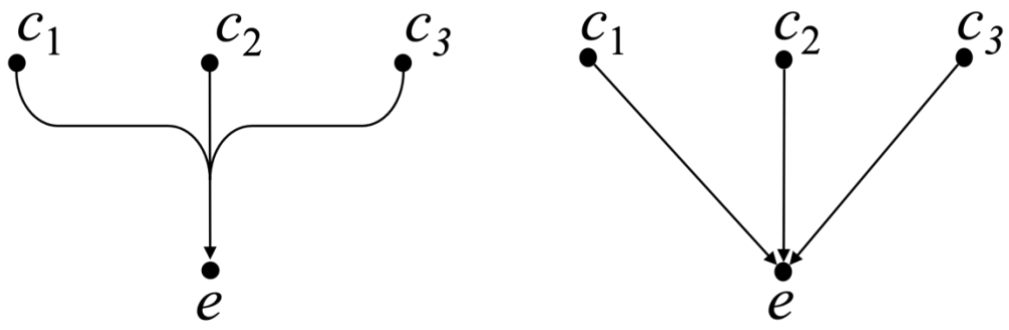
specific causal relationships between *properties or kinds* (*C* causes *E*) (Hausman, 2005). The key differences between type and token relationships can be seen in probabilistic contexts in which a token event, *c*, may be known as having a *tendency* to cause a token outcome event, *e*, by raising its probability, which would establish *type* causality between the class C and E. In this way, type causation refers to generalizations of actual or possible relationships, whereas token causation occurs only if both cause and effect occur. The benchmark we introduce here concerns reasoning about token causality.

Numerous accounts have been proposed to explain the conditions under which one thing is considered the cause of another. A view of causality aimed at describing how events *become* causal was bolstered in the 40s by Albert Michotte; whose experiments involved visual stimuli depicting what appeared to be one shape moving towards and making contact with a previously static shape, which is then "launched" into motion along the same direction as the first (Michotte, 1946). These simple launching effects remain a powerful abstraction for studying the perception of causality and laid the groundwork for process theories of causation. According to process accounts, some candidate, *c*, may be considered a cause of an outcome, *e*, if they are connected via the physical transfer of some quantity or amount that moves from *c* to *e* (Phil Dowe, 1992; Salmon, 1994; Wolff, 2007).

An alternative approach was motivated by the desire to give a satisfying account of the semantics of statements involving causal claims. It argued that causality is conditional on the co-occurrence of events such that *c* is deemed a cause of *e* if the occurrence of *e* depends on the occurrence of *c*. This dependence can be established if the occurrence of *c* raises the probability of *e*, or through *counterfactual dependence* where *e* does not occur in any possible counterfactual in which *c* does not occur (D. Lewis, 1979; Suppes, 1968). More recent models to describe human causal reasoning have combined the strengths of process and dependence theories together and include criteria for both physical force transfer as well as counterfactual dependence (Gerstenberg et al., 2021).

Although early proposals to define causality, such as those offered by process and dependence theories, provide a useful theoretical foundation, the dataset described below focuses, instead, on exploring how *people* intuitively reason about the causes of events. Reaching a unified account of human

causal judgments is complicated, in part, because a mapping from cause to effect is rarely, if ever, one-to-one. Causal *selection* is the process by which humans decide on the definitive single cause of an event with multiple necessary or sufficient antecedent conditions. Most events in real life occur through the confluence or consequence of many distinct causal conditions. Accordingly, humans can attribute causal strength across a set of candidate events in a graded fashion (Danks, 2017; Halpern & Hitchcock, 2013). When considering which events might be causes, the various links within and across causal conditions to their effects define the structure of a causal system. The causal selection benchmark dataset focuses on causal attribution judgments for singular token outcome events in causal systems across various structural configurations (Figure 4.1). In particular, we focus on the influence of normality on causal selection because there exists a clear opportunity to systematically fill in the gaps between a binary construction of normality as categorically rare or common. In what follows, we describe the effect of normality on human causal judgments as it has been studied thus far and outline the methods and manipulations we used to more systematically demonstrate these effects in the benchmark dataset.



A. **B.**

Figure 4.1, Directed acyclical graphs depicting the structure causal relationships. A. Conjunctive structure. B. Disjunctive structure

4.1.2 The influence of normality on causal attribution

Effect of normality on causal attribution in conjunctive systems

A common structure used to study causal selection involves token outcomes that result from the conjunction of two antecedent causal events. In these structures, neither antecedent causal event, alone, is sufficient to bring about the outcome. However, the *conjunction* of both events, together, is necessary. One instantiation of conjunctive causal structure comes from Icard, et al (2017):

> Prof Smith works at a large university. At this university, in order to get new computers from the university, faculty, like Professor Smith, must send an application to two administrative committees for approval, the IT committee, and the department budget committee. Professor Smith will be able to get her new computers if the IT committee approves her application AND the department budget committee approves her application. Both committees must approve the application for her to get the new computers.
>
> Prof Smith sends in her applications. Each committee meets independently and they decide without talking to each other, but their meetings are scheduled for the exact same time. The IT committee approves her application and the department budget committee approves her application. So Prof Smith got her new computers.

The outcome in the example above counterfactually depends on the conjunction of both candidate causes occurring (i.e. if one committee approved the request, but the other committee did not, Professor Smith would not have received new computers). Therefore people may attribute responsibility for Professor Smith's new computers equally to both committees (Zultan et al., 2012). But consider the following information about the normality of each necessary cause:

The IT committee almost always approves these applications. The department budget committee almost never approves these applications. The budget committee is notorious for turning down almost every application they receive.

Given this information about the relative normality of both necessary causal conditions occurring, if the outcome remains that Professor Smith received new computers, people consistently attribute more responsibility to the budget committee's surprising approval. This effect is referred to as *abnormal inflation*: the responsibility attributed to one cause in a conjunctive causal system increases as a function of its abnormality relative to other necessary causes (Hart & Honoré, 1985; Hilton & Slugoski, 1986; Icard et al., 2017; Kahneman & Miller, 1986; Kahneman et al., 1982). Now consider how responsible the IT committee is for Professor Smith receiving new computers. People tend to attribute less responsibility to the IT committee in this case since it is seen are relatively more probable than the other necessary antecedents. This pattern is known as *supersession*, wherein the responsibility attributed to one cause in a conjunctive system decreases as a function of its normality relative to other necessary causes (Kominsky et al., 2015).

Effect of normality on causal attribution in disjunctive systems

Another common structure used to study causal selection involves token outcomes that result from the disjunction of two antecedent causal events. In these structures, either antecedent causal event, alone, is sufficient to bring about the outcome. Therefore, one or the other antecedent event could, individually, cause an outcome. Returning to the example from Icard, et al (2017) in a disjunctive scenario:

Prof Smith works at a large university. At this university, in order to get new computers from the university, faculty, like Professor Smith must send an application to two administrative committees for approval, the IT committee, and the department budget committee. Prof Smith will be able to get her new computers if the IT committee approves her application OR the

department budget committee approves her application. Only one of the committees needs to approve her application for her to get new computers.

Prof Smith sends in her applications. Each committee meets independently and they decide without talking to each other, but their meetings are scheduled for the exact same time. The IT committee approves her application and the department budget committee approves her application. So Prof Smith got her new computers.

The outcome in the example above now counterfactually depends on only one of the two causal events occurring (i.e. if one committee approved the request, but the other committee did not, Professor Smith would still have received new computers). Therefore people may attribute responsibility for Professor Smith's new computers equally to both committees. But consider, again, the following information about the normality of each sufficient cause:

The IT committee almost always approves these applications. The department budget committee almost never approves these applications. The budget committee is notorious for turning down almost every application they receive.

Given this information about the relative normality of both sufficient causal conditions occurring, if the outcome remains that Professor Smith received new computers, people consistently attribute less responsibility to the budget committee's surprising approval. This effect is referred to as *abnormal deflation*: the responsibility attributed to one cause in a disjunctive causal system decreases as a function of its abnormality relative to other sufficient causes (Icard et al., 2017). Now consider how responsible the IT committee is for Professor Smith receiving new computers. The responsibility people assign to the IT committee, in this case, remains unchanged, despite it being relatively more normal than another individually sufficient antecedent, the budget committee's approval. This *reverse supersession* is

observed, wherein the responsibility attributed to one cause in a disjunctive causal system may stay the same or increase as a function of its normality relative to other causes (Kominsky et al., 2015).

4.1.3 Mixed Causal Structures

Pure two-variable conjunctive and disjunctive structures describe simplifications of, otherwise, extremely high-dimensional systems. However, very little work has investigated more complex structures that combine conjunctive and disjunctive causal conditions for a single outcome. Our benchmark dataset creates opportunities for researchers to test existing models in novel contexts as well as develop new theoretical accounts of causal attribution in mixed causal structures described below.

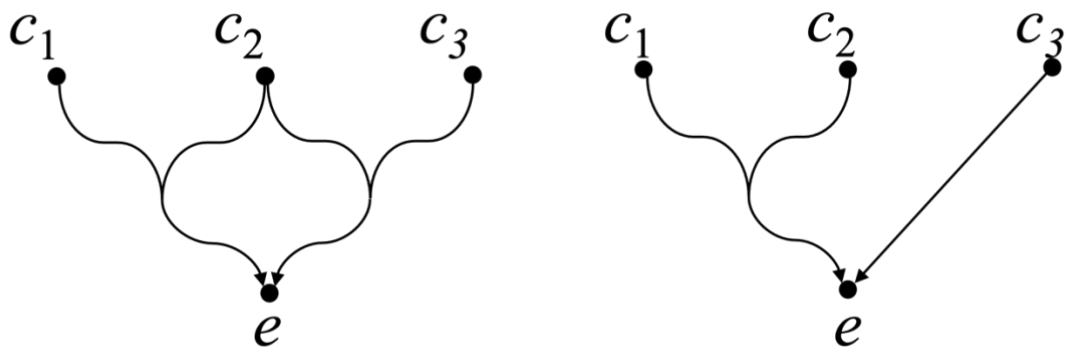
Mixed conjunctive systems

In a mixed conjunctive system, a disjunctive causal structure is embedded within a conjunctive one such that two events must still occur to bring about the token outcome, one of which, however, occurs through the disjunction of two distinct events. For example, imagine a budding technology company looking to hire a new senior data scientist for their team. The job listing describes the requirements needed for consideration by the hiring manager. In order to be qualified for the job, the listing states that applicants must have a PhD *and* pass a coding assessment test in either one of two programming languages, Python or Javascript. Considering why a particular applicant was offered the position requires causal selection in a mixed conjunctive system. An applicant was offered a job, e, because they have a PhD, c_1, *and* they've passed the coding assessment in Python *or* Javascript, c_2. In other words, e can only occur by the conjunction of c_1 and c_2, where c_2 results from a *dis*junction of two individually sufficient events (Figure 4.2A). We test causal attribution ratings in this mixed conjunction causal structure.

Mixed disjunctive systems

The causal selection benchmark dataset also contains causal judgments in mixed disjunctive structures. More specifically, in a mixed disjunctive system, a conjunctive causal structure is embedded within a disjunctive one such that two events are individually sufficient to bring about the token outcome,

one of which, however, occurs through the conjunction of two distinct events. Consider that the same company is hiring for a different role, that of User Experience Researcher. The job posting for this position describes the requirements needed for consideration by the hiring manager. In order to be qualified for the role, the job listing states that applicants must have a PhD *or* pass a coding assessment test in two programming languages, Python *and* Javascript. Considering why a particular applicant was offered the position requires causal selection in a mixed disjunctive system. An applicant was offered a job, *e*, because they have a PhD, c_1, *or* they pass the coding assessment for Python *and* Javascript, c_2. In other words, *e* can occur by the disjunction of c_1 or c_2, where c_2 is the result of a *con*junction of two events (Figure 4.2B). We also test causal attribution ratings in this mixed disjunctive causal structure.



A B.

Figure 4.2, Directed acyclical graphs depicting the mixed structure causal relationships. A. Mixed conjunctive structure where the conjunction of c2 with c1 or c3 is necessary for the outcome. B. Mixed Disjunctive structure where the conjunction of c1 and c2 or the occurrence of c3, would both be sufficient to cause the outcome.

4.1.4 The current contribution

The causal selection benchmark dataset improves upon existing work in three major ways. First, we extend the causal structures mentioned above to include up to three conjunctive or disjunctive causes of a single outcome event. These augmented causal systems provide an opportunity to explore the gradation of causal attribution when more variables are considered in the reasoning process. Second, we parametrically modulate the influence of normality for all three causal events across a broad range of quantitative values. Covering such a vast space of possibilities will allow researchers to explore how causal judgments change in proportion to systematic changes in the normality of an event. Finally, we examine causal attribution in more complicated mixed causal structures that come closer to the complexity of the causal judgments in the real world. These mixed structures provide novel contexts to test existing proposals or develop entirely new computational models of causal selection.

4.2 Benchmark Data Collection Methods

4.2.1 Participants

All participants included in the benchmark dataset were recruited from Prolific (www.prolific.co) in the period from November-December 2022. Inclusion criteria consisted of English fluency and an approval rating for data submissions on Prolific of > 90%. Participant demographic data are provided by Prolific in comma-separated value format. Basic demographic descriptives are provided in *Table 4.1*.

Group	N	Age	Sex
Pure conjunctive	399	*M* = 32.63, SD = 11.77	139 females
Pure disjunctive	410	M = 30.06. SD = 12.09	205 females
Mixed conjunctive	393	M = 28.17, SD = 9.92	195 females
Mixed disjunctive	397	M = 31.13, SD = 11.97	195 females

Table 4.1, Sample sizes and participant demographics in each group included in the causal selection benchmark

4.2.2 Causal System Manipulation

The structure of the causal system being observed by each participant was manipulated between groups through the instructions of the task. Participants were informed about the rules for winning a lottery, which varied according to the participant group. The instructions presented to participants across the four causal structures probed in the dataset can be found in *Table 2*.

Group	Task instructions
	"In this study, you will observe a series of winning players' lottery outcomes. The rules of the lottery are as follows: A player draws one ball from each of three different jars, A, B, & C. In order to win the lottery…"
Pure conjunctive	"...a player must draw three red balls - one from each of the three jars. So if they draw three red balls, they will win."
Pure disjunctive	"...a player must draw at least one red ball from the three jars. So if they draw one or more red balls, they will win."
Mixed conjunctive	"...a player must draw a red ball from jar A and must also draw at least one red ball from jar B or C. So if they draw a red ball from jar A, and one or more red balls from jar B and C, they will win."
Mixed disjunctive	"... a player must draw a red ball from jar A or must draw a red ball from both jar B and C. So if they draw a red from jar A or red balls from both jar B and C, they will win. "

Table 4.2, Task instructions in each of the four different causal systems presented between participants in the causal selection benchmark dataset

4.2.3 Stimuli and Procedure

The stimulus for all experiments was presented through a custom web application built using the Flask python micro-framework. The application was hosted on a custom-built, encrypted server (https://www.linode.com/), secured with a firewall and other security precautions. Connections to the experiment server were screened to ensure that the experiment was not accessed through a virtual privacy network, proxy, relay network, or tor node. Before beginning the experiment, authenticity validation was required using Google's reCAPTCHA v2. All participants were assigned a unique 16-character, study-specific ID, and data was written to an encrypted SQLite relational database.

After providing informed consent, subjects were instructed that they would be observing a series of winning lottery outcomes. We explained that to play the lottery, players needed to draw a ball from three different jars, which each contained some amount of red and blue balls. Participants were informed about the rules for winning the lottery, which varied according to the participant group as described above and presented in *Table 2*. After providing instructions that manipulate the causal structure being observed, we showed participants examples of both winning and losing draws, according to the causal structure being observed. Importantly, all other aspects of the experiment remained unchanged between groups. We then explained that after observing the lottery outcome, they would be presented with the percentages of red and blue balls contained in each jar, signifying the player's chances of drawing either color. At the same time, the actual draw that occurred from each jar is also presented (always a red ball). Participants were instructed to use sliders on the screen to indicate the extent to which they judged each jar causally responsible for the lottery outcome. Participants completed a brief comprehension check to ensure they fully understood the task instructions and the lottery's causal structure. The comprehension check asked participants to select possible draw scenarios that would lead to a win given the rules of the lottery. Participants were redirected back to the instructions if they did not answer the comprehension check correctly. As part of the comprehension check, we also asked participants a free text-response question, "What information is shown about each jar?". The answer to the free text-response question was not

checked for accuracy online during the experiment, but in the vast majority of cases, participants correctly entered some version of the response "*The chances of drawing a red or blue ball.*"

All participants completed ten trials. On each trial, subjects first saw an image of cash with the word "Win!" written, signifying that the lottery outcome they were observing came from a player who met the necessary and sufficient conditions to cause the outcome event of winning the lottery. Next, the percentages of red and blue balls contained in each jar were displayed, as well as a gif depicting a red ball drawn from each glass jar containing red and blue balls in proportion to the percentages displayed (each gif looped once and remained static on the final frame depicting a red ball above the jar). All images and gifs presented to participants were made using Blender v2.9 3D computer graphics software. Importantly, the color drawn from each jar was the same across jars within a trial and across each causal structure such that the only difference between each causal structure group condition was the instructions given to participants outlining the conditions that were necessary and sufficient for a person to win the lottery (*Table 2*). Finally, three slider bars appeared next to each jar where participants indicated the extent to which they agreed with the statement, "This player won the lottery because they drew a red ball from jar [A/B/C]." Slider values were initialized halfway between the range from 0 ("Completely disagree") to 100 ("Completely Agree"). The numeric value of the slider was displayed and updated as participants selected their answers (*Figure 3.2*). After each trial, participants were asked to indicate which of the two colored balls was more likely to be drawn from one of the three jars, randomly selected from the preceding trial. After submitting their answer, subjects were given feedback on whether they answered correctly before moving to the next trial.

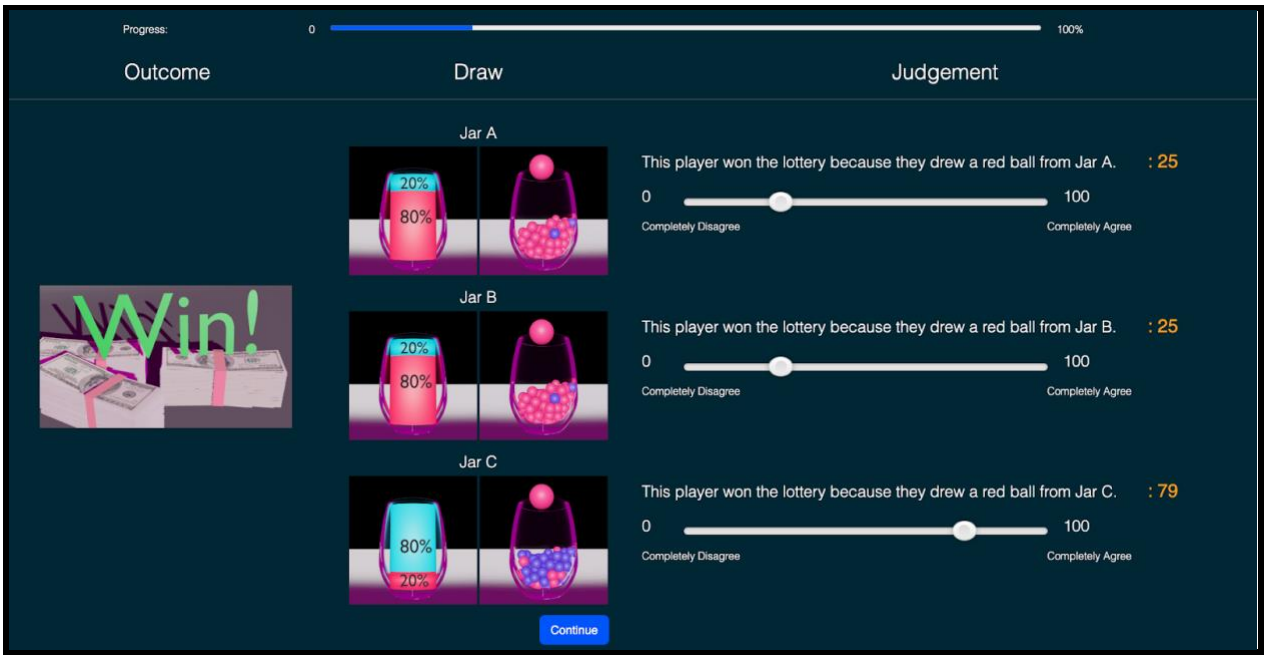


Figure 4.3, Trial display viewed by participants in all conditions. First, the outcome event was shown, followed by the percentage of red and blue balls and the ball drawn from each jar (always red). Finally, subjects moved a slider to indicate how causal each jar was for the winning outcome

4.2.4 Normality Manipulation

The probability of drawing a red ball from each jar, $P(c_i = 1)$, on a given trial was drawn at random from the set of 125 ordered triplets, $\mathbb{J}^3$, spanning the tertiary Cartesian product of the set {.2, .4, .6, .8, 1}. In other words, we included all possible combinations of the probability of drawing a red ball for each of the three jars. For each of the four causal structures tested, each point in this space was sampled an average of 32.208 times ($M_{\text{conjunction}}$ = 32.07, $SD_{\text{conjunction}}$ = 5.66; $M_{\text{disjunction}}$ = 32.87, $SD_{\text{disjunction}}$ = 5.43; $M_{\text{mixed conjunction}}$ = 31.80, $SD_{\text{mixed conjunction}}$ = 5.72; $M_{\text{mixed disjunction}}$ = 32.08, $SD_{\text{mixed disjunction}}$ = 5.26; 3). Across all four causal structures we tested, each triplet was sampled an average of 128.83 times (SD=10.91, Figure 4.4).

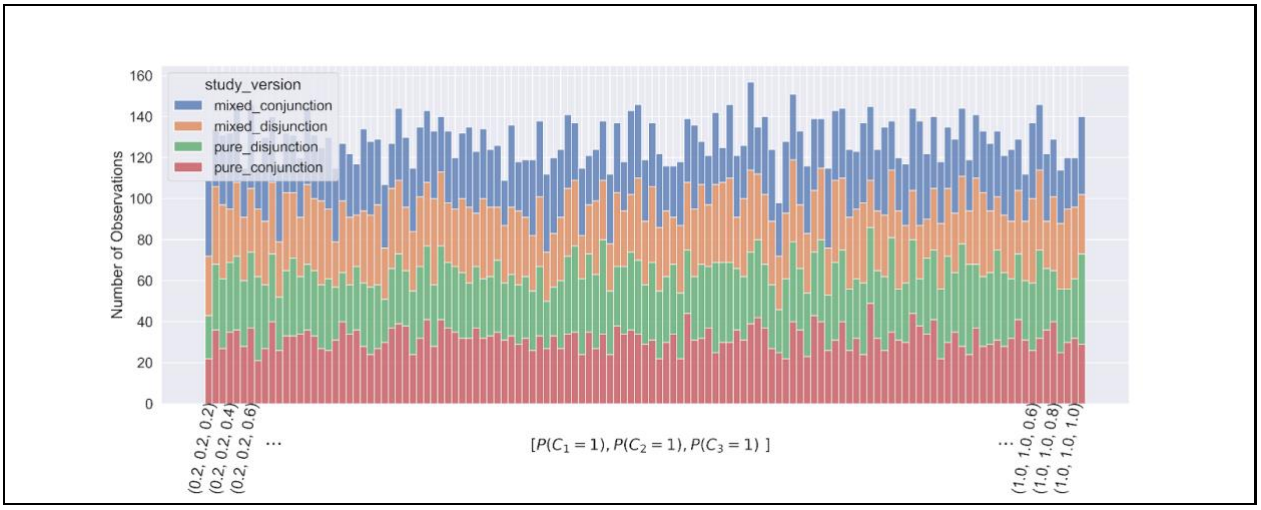


Figure 4.4, Distribution of observations for each of the 125 levels of the normality condition across the four causal structures tested.

4.3 Data

The SQL DBLite database for the causal selection benchmark contains two relational tables. The *Subjects* table contains one entry per participant. Each row includes boolean columns representing whether the participant 1.) returned successful verification by Google's reCAPTCHA API, 2.) provided informed consent, 3.) passed the instruction comprehension check, and 4.) completed the experiment. Each row also includes two SQL DateTime columns indicating start and completion times as Python DateTime objects recorded on consent submission and debrief presentation respectively. Finally, each row of the Subjects table contains three string columns indicating 1.) the participants' answer to the free text-response question asked during the instruction comprehension check, 2.) the type of causal structure observed across all trials by the participant, and 3.) the participant's unique, 16-character alphanumeric study ID.

The second table in the benchmark database, *Trials*, contains ten entries per participant; one for each trial completed. Each row/trial contains a column indicating 1.) the participant's alphanumeric study ID; 2.) the trial number; 3.) a column indicating the probability of a given causal event, $P(c_i = 1)$, for each of the three causal events under consideration; 4.) for each causal event (c_1, c_2, c_3), a column containing a list of integers indicating the value after each move of the slider. The final value in this list indicates the participant's final attribution rating of a given causal event submitted for that trial. Each row of the *Trials* table also contains information about which randomly chosen causal event was probed after the trial, as well as a boolean indicating whether or not the participant correctly recalled if the event was more or less likely to occur in that trial. These SQL tables are joined and provided as comma-separated values for convenience. All data and materials will be made available at https://github.com/BryanGonzalez262/ChooseWhy.

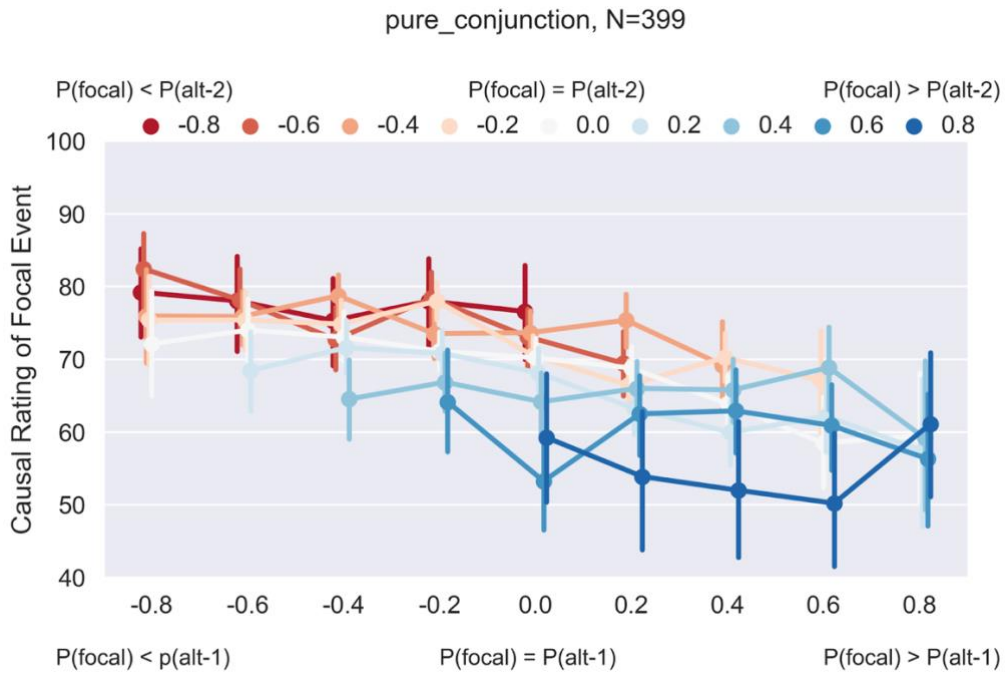
4.3.1 Data Quality and preliminary trends

On average, participants completed the task in ~13.7 minutes with no differences in duration times across participant groups. Recall that after each trial, a randomly chosen jar was used in an attention check in which we asked participants whether a red or blue ball was more likely to be drawn from the selected jar. The mean accuracy of these attention checks across all participants was 94.4%.

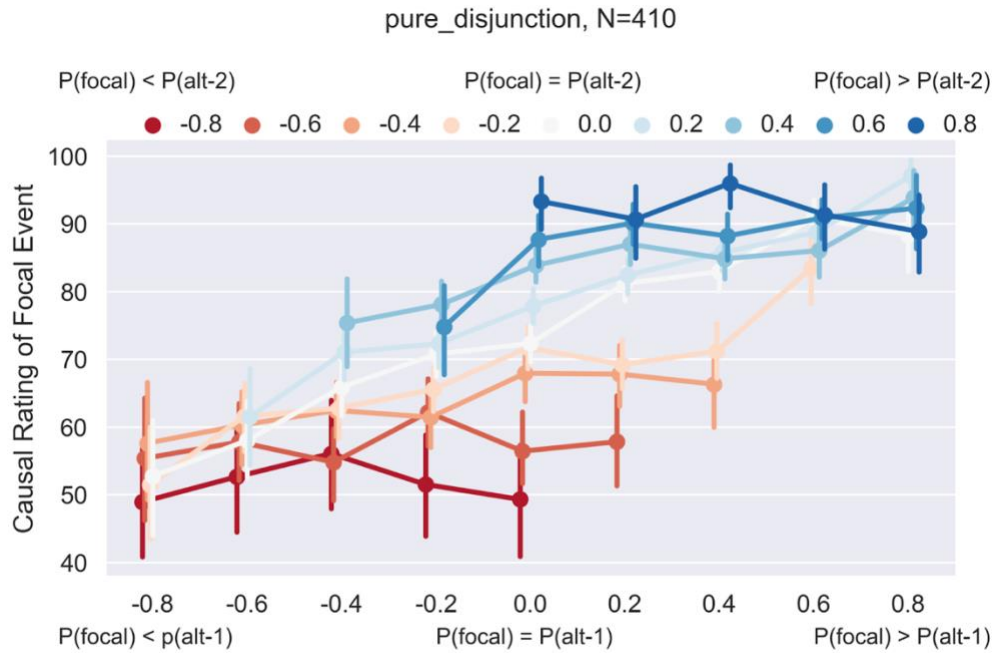
Finally, we plotted causal ratings against relative probability values to determine if the effects of normality found in the literature were replicated in the paradigm we use here. *Figure 4.5* illustrates the graded influence of normality on the casual judgments of a single, "focal", event. In other words, causal judgments for one jar (the focal event), as a function of its normality relative to each alternative jar. The x-axes of *Figure 4.5* represents the relative normality of the focal event against the first alternative, i.e the signed difference between the probability of drawing a red ball from one jar and the probability of drawing a red ball from the first alternative jar. Furthermore, the hue gradient of *Figure 4.5* represents the relative normality of the focal event against the second alternative, i.e the signed difference between the

probability of drawing a red ball from one jar and the probability of drawing a red ball from the second alternative jar.

In the "Pure" Conjunction group (*Figure 4.5,* A.), abnormal inflation against the *first* alternative jar (x-axis) can be seen from the *increasing* causal judgments moving across the x-axis from the center, *P(focal jar) = P(alternative jar 1)*, to the left. *P(focal jar) < P(alternative jar 1).* Furthermore, abnormal deflation against the first alternative jar (x-axis) can be seen from the *decreasing* causal judgments moving across the x-axis from the center, *P(focal jar) = P(alternative jar 1)*, to the right. *P(focal jar) > P(alternative jar 1).* Also in the "Pure" Conjunction group (*Figure 4.5,* A.), abnormal inflation against the *second* alternative jar (hue) can be seen from the *increasing* causal judgments moving across the hue gradient from the white, *P(focal jar) = P(alternative jar 2)*, to the red. *P(focal jar) < P(alternative jar 2).* Furthermore, abnormal deflation against the second alternative jar (hue) can be seen in *decreasing* causal judgments moving across the hue gradient from the white, *P(focal jar) = P(alternative jar 2)*, to the blue. *P(focal jar) < P(alternative jar 2).* Note that more complex patterns emerge at the extremes of each alternative (-0.8 and 0.8) whether the alternative event (-0.8) or the focal event (0.8) are certain to occur with a probability of 1. The general patterns observed in the "Pure" Conjunction group appear to reverse in the "Pure" Disjunctive case (*Figure 4.5,* B.).



A.



B.

Figure 4.5, Causal judgments of a focal event (jar) as a function of its normality relative to both alternative causal events. The x-axis represents the relative normality of the focal event as compared to the first alternative (P(focal) - P(alternative-1)). Hues represent the relative normality of the focal event as compared to the second alternative (P(focal) – P(alternative-2)). A. "Pure" 3-jar conjunction. B. "Pure" 3-jar disjunction.

4.4 Discussion

Definitively characterizing what makes an event the "cause" of a singular outcome has remained elusive (Hume, 1748), leading to various hypotheses of causal attribution that often make diverging predictions of how humans intuitively judge one event as the cause of another (Halpern & Hitchcock, 2013). One possible reason for such heterogeneity of accounts is the oversimplification of factors known to play a role in causal selection.

This causal selection benchmark dataset augments the study of human causal attribution in multiple ways. Existing work on human judgments in causal systems typically restricts investigation to elementary systems in which just two events are considered to jointly or individually cause an outcome. In these contexts, reasoners often pit causal events against one another, choosing one as *the* cause of an outcome. However, the distinction between causes and non-causes need not be dichotomous. In fact, human attribution of causality is often graded in nature, where multiple events are considered causes of the same outcome to varying extents. The benchmark dataset described here captures causal attribution judgments in systems involving dynamic relationships across a larger set of three distinct events. This simple, yet elegant modification will allow researchers to examine how humans diffuse responsibility for an outcome among a larger set of causes.

This benchmark dataset also precisely manipulates the powerful influence of normality on the causal selection process (Cushman et al., 2008; Hitchcock & Knobe, 2009). With the exception of Morris et al (2019), most of the prior efforts to elucidate how our expectations of events impact the causality we attribute to them have operationalized this variable in broadly qualitative ways. However, this coarse approach produces more questions than it answers. Is the relationship between norms and causal attribution linear? What defines the threshold beyond which an event is considered 'abnormal'? Our benchmark dataset is the first of its kind to systematically manipulate the normality of events in a causal system across a continuous 3-dimensional space of probabilities. Exploring normality in this quantitative way presents a new opportunity to study the relationship between normality and causal judgments using

more comparable numerical scales. Furthermore, since the effect of normality has been shown to hold across both descriptive and prescriptive norms (Bear & Knobe, 2017; Icard et al., 2017), insights gained from the quantitative manipulations employed here may have broader implications for the effect of normality more broadly.

Finally, this benchmark dataset captures causal attribution judgments in novel configurations of causal systems that have been unexplored until now. This offers the chance to test the robustness of existing models, as well as invite new ideas that expand how researchers across disciplines characterize the causal selection process.

4.4.1 Avenues for future research

The causal selection benchmark dataset and materials were designed to allow simple expansions in a number of important directions. Below we describe additional variables for consideration, how these variables have been shown to interact with normality, and how our paradigm can be adjusted to study the effect of these factors on causal selection.

Omission

Barry promises to water Alice's plants while she is on vacation. Suppose that while Alice is away, Barry completely forgets to water Alice's plants. Upon returning from her trip, Alice finds that her plants have tragically died. Clearly, Barry's failure to water Alice's plants is the cause of their death. The plants dying counterfactually depend on this omission (i.e. if Barry had watered the plants, they would not have died.). In fact, any event counterfactually depends on the omission of other events that would otherwise prevent it from occurring. This notion exposes an endless amount of necessary antecedents that, by virtue of *not* happening, lead to a token outcome (McGrath, 2005; Menzies, 2004; Wolff et al., 2010). Countless omissive events also qualify as conjunctive causes for Alice's dead plants. Beyonce not watering Alice's plants is also a causal omission since if Beyonce had watered Alice's plants, they would not have died. Nonetheless, most people would blame Barry, and not Beyonce, for Alice's dead plants in

this conjunctive system. How people solve this variable selection problem in this scenario involves the relevance of counterfactuals given what is considered normal (Henne et al., 2017).

One can imagine a different scenario in which Alice's plants survive after Barry remembers to water them. Although not explicit the conjunction of an omission and commission are each necessary for this outcome. That is, Barry watering Alice's plants and the omission of some preventative event (e.g. Barry is not called to work for an emergency meeting) are both necessary conditions for Alice's plants to survive. In this scenario, people consistently judge the commissive event of Barry watering the plants as more responsible for the outcome than the omissive one in which Barry was not called to work. In causal systems requiring the conjunction of an omissive and commissive event, humans consistently attribute more responsibility to the commissive event across various contexts (Cushman & Young, 2011; Spranca et al., 1991; Walsh & Sloman, 2011; Yeung et al., 2022). Again, the pervasive influence of normality accounts for why causality is attributed to Barry's commissive actions, rather than the necessary commissive event (Schaffer, 2005). In other words, people's causal judgments of omissive events are influenced by their expectations, which dictate the most relevant omissions to test for counterfactual dependence (Gerstenberg & Stephan, 2021). Other work has demonstrated that the effect of normality reverses in disjunctive systems involving commissive or omissive events (Henne et al., 2019). Further research is needed to explore the influence of expectations more methodically using quantitative manipulations like the ones we employ in the causal selection benchmark.

The paradigm described in the methods section of this report can be easily amended to study the effects of normality on causal selection judgments of omissive events. Simply modifying the written rules of the lottery, described in section 4.2.2, and token events to reflect causal conditions in which *not* drawing a red ball leads to a winning lottery outcome, will allow interested researchers to explore the role of normality on judgments of causation by omission.

Value

Another factor we believe is ripe for investigation is the influence of value on causal selection. The value associated with token events and their effects has, so far, demonstrated intriguing influences on how people select causes.

Value has often been operationalized in vignettes that describe agents as causing an outcome by violating a moral norm. In these situations, there is a link between the valence of a necessary cause and the valence of the outcome. In one classic example, participants read the description of a person rushing to get who got into an accident. The accident was caused by the conjunction of two necessary events: the driver was speeding and it was raining. The authors manipulated valence by telling participants the person was rushing for a bad (to hide cocaine) or good (to hide an anniversary gift) reason. This manipulation lead to systematic changes in causal selection such that the rain was just more responsible for the accident if the driver was good, but speeding was deemed more causal if the driver was bad. (Alicke, 1992). Fascinating questions remain open about whether moral prescriptive norm violations exert an equal or stronger influence than statistical descriptive norm violations on causal selection judgments.

The paradigm described in the methods section of this report can be easily amended to study the effects of normality on causal selection judgments of valenced events. Simply modifying the written rules of the lottery, described in section 4.2.2, and token events to reflect causal conditions in which drawing a red ball is associated with positive or negative consequences independent of the lottery outcome, will allow interested researchers to explore the role of normality on judgments of valenced causal events.

4.5 Conclusion

The process by which we attribute causal responsibility for specific outcomes to specific antecedent events is known as token causal selection (Halpern & Hitchcock, 2013; Hausman, 2005). Prior work has described causal events as categorically common or uncommon and demonstrates that our

expectations of an event's normativity interact with our knowledge of causal relationships to influence what we select as the cause of an outcome (Icard et al., 2017). We introduced a publicly available dataset structured around four novel configurations of causal relationships. Furthermore, we quantitatively manipulate the influence of normality to systematically explore the continuous space of an event's probability from unlikely to certain. This large sample of human judgments bears out some of the known influences of normality such as the tendency to consider relatively abnormal events as more causal in conjunctive cases, and relatively normal events as more causal in disjunctive cases. Taken together, our benchmark dataset may serve researchers interested in causal selection as a growing testbed for diverging theoretical and computational models proposing to characterize how humans select causes.

Chapter 5.

"Those who have a '*why'* to live, can bear almost any '*how*'."

- Friedrich Nietzche x Viktor E. Frankl

5.1 General Discussion

This dissertation has examined epistemic mental state attribution, the interaction of social perception and causal reasoning, and the more domain-general influence of our normative expectations on the process of selecting causes of events.

Outside of controlled psychology experiments, we mentalize about what others might know to be true much more often than what they falsely believe (Phillips & Norby, 2019). The first aim of this dissertation was to examine whether our capacity to represent another agent's knowledge results from more fundamentally basic processes than those recruited to represent their beliefs. I demonstrate that evaluations of other people's knowledge occurs faster than corresponding evaluations of their beliefs. Across five experiments I excluded the possibility that this effect results from the linguistic idiosyncrasies of our specific stimuli, or the English language thereof. I also demonstrated that the relative speed advantage in evaluating knowledge ascriptions generalizes to the larger class of factive and non-factive mental states and is instantiated in reduced neural responses during factive as compared non-factive mentalizing in a brain region known to be recruited when thinking about others. These results imply that we can represent the propositional knowledge of others without the need to further represent their simultaneous beliefs therein. This suggests that representations that are consistent with our existing sense

of reality are fundamentally easier to hold than those that may run against it. My conclusions speak to the broader function of our mentalizing abilities to teach us about the world beyond our own experience of it, extending our senses outwards through the eyes of others.

The path from observing others' external behaviors to inferring their latent internal mental states rests on the crucial assumption that other agents choose actions in order to maximize their subjective well-being (Dennett, 1983; Jara-Ettinger et al., 2016). As such, the inferences we make about an agent's mind are over the subjective preferences and epistemic states that are causally responsible for an agent's actions. Thus, the process of mentalizing is inextricable from that of causal reasoning more broadly (Premack & Woodruff, 1978, p. 525). The second aim of this dissertation is to examine how the perception of agency and prescriptive social norms interact to influence our intuitions of how agents and objects cause events in the world. Using anthropomorphic stimuli, I find evidence that agents are not judged as more causal to an outcome than objects by virtue of simply appearing "alive" and goal-directed. Instead, I argued that the differences in causal attributions to agents and objects derive from the effect of equifinality inherent to agency. To get a sense for how this matters, consider an example from William James in which he compares the futures of intentional agents and inanimate iron filings (Lombrozo, 2010, p. 309):

> Romeo wants Juliet as the filings want the magnet; and if no obstacles intervene he moves towards her by as straight a line as they. But Romeo and Juliet, if a wall be built between them, do not remain idiotically pressing their faces against its opposite sides like the magnet and the filings [when a card is placed between them]. Romeo soon finds a circuitous way, by scaling the wall or otherwise, of touching Juliet's lips directly. With the filings the path is fixed; whether it reaches the end depends on accidents. With the lover it is the end which is fixed, the path may be modified indefinitely (James, n.d., p. 20) .

This description of equifinality suggests that outcomes brought about by agents can be changed only through a change in the agent's goal, rather than a change in the actions by which a 'fixed end' is pursued. In other words, merely simulating variations in the actions of goal-directed agents are unlikely to produce the counterfactual contrasts needed for causal claims if the agent's goal remains fixed and is nonetheless achieved despite the perturbance. Accordingly, the counterfactuals we consider for agents involve alternative intentions. However, the ability to change the future only tells us that alternatives are possible for agents that cause events that are not for possible causal objects. It is our normative expectations that tell us which alternatives are most probable.

A step toward understanding *why* we think something occurs is to consider *how* information about causal relationships shapes what we consider to be a cause at all. Consistent patterns have emerged in the factors influencing our causal attributions (Cushman & Young, 2011; Hitchcock & Knobe, 2009). However, our understanding of these influences can be improved by examining their effects more systematically. In chapter 4 of this dissertation, I focused on the effect of statistical normality on the way in which humans reason about the causes of events. I introduced a large, publicly available benchmark dataset of causal selection judgments I hope serves as a testbed for existing and new theories of causal cognition. This benchmark holds immense potential for researchers to closely examine the gradation of responsibility judgments over multiple causal events. Furthermore, the systematic manipulation of the normality of these causal events provides the unprecedented opportunity to adjudicate competing theories of how we answer the question: 'why?'.

5.2 Mentalizing as Causal Inference

Mental causation, or the capacity of mental states to cause physical events, is a controversial problem in philosophy of mind. Many psychologists take for granted, however, the notion that mental states, and the complexities therein, are the proximal causes of volitional behavior. As such, the process

of inferring another agent's mental state is really one of selecting the latent cause of the past or future behavior. Researchers interested in characterizing our mentalizing abilities have much to gain by harnessing the insights known or theorized about the domain-general mechanisms of causal reasoning.

References

Adolphs, R. (2009). The social brain: neural basis of social knowledge. *Annual Review of Psychology*, *60*, 693–716. https://doi.org/10.1146/annurev.psych.60.110707.163514

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, *63*(3), 368–378. https://doi.org/10.1037/0022-3514.63.3.368

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, Norm Violation, and Culpable Control. *The Journal of Philosophy*, *108*(12), 670–696. https://doi.org/10.5840/jphil20111081238

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward Brief "Red Flags" for Autism Screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist in 1,000 Cases and 3,000 Controls. *Journal of the American Academy of Child and Adolescent Psychiatry*, *51*(2), 202-212.e7. https://doi.org/10.1016/j.jaac.2011.11.003

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953–970. https://doi.org/10.1037/a0016923

Aronson, J. (1971). On the Grammar of "Cause." *Synthese*, *22*, 414–430.

Baron - Cohen, S. (1989). Perceptual role taking and protodeclarative pointing in autism. *The British Journal of Developmental Psychology*, *7*(7), 113–127. https://doi.org/10.1111/j.2044-835X.1989.tb00793.x

Baron-Cohen, S. (1997). *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press. https://play.google.com/store/books/details?id=MDbcNu9zYZAC

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind" ? *Cognition*, *21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. https://books.google.com/books?hl=en&lr=&id=GZkmKBY664kC&oi=fnd&pg=PR9&dq=Barts

ch+K+Wellman+H+M+(1995)+Children+talk+about+the+mind+Oxford+university+pr&ots=Hpok8gjWv5&sig=iKy5ArC4lzKDvhyAalJGRO_tRh0

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition*, *167*, 25–37. https://doi.org/10.1016/j.cognition.2016.10.024

Beebee, H. (2004). Causing and nothingness. *Causation and Counterfactuals*, *291308*. https://books.google.com/books?hl=en&lr=&id=JJL3xRNrCpUC&oi=fnd&pg=PA291&dq=beebee+2004&ots=_fYyA_GlVl&sig=DzUHdM30Te06egPBKAX6YLXI_S0

Bennett, J. (1978). Some remarks about concepts. *The Behavioral and Brain Sciences*, *1*(4), 557–560. https://doi.org/10.1017/S0140525X00076573

Bernstein, S. (2015). The metaphysics of omissions. *Philosophy Compass*, *10*(3), 208–218. https://doi.org/10.1111/phc3.12206

Bishop, J. M. (2020). Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It. *Frontiers in Psychology*, *11*, 513474. https://doi.org/10.3389/fpsyg.2020.513474

Bräuer, J., Call, J., & Tomasello, M. (2007). Chimpanzees really know what others can see in a competitive situation. *Animal Cognition*, *10*(4), 439–448. https://doi.org/10.1007/s10071-007-0088-1

Bricker, A. M. (2020). The neural and cognitive mechanisms of knowledge attribution: An EEG study. In *Cognition* (Vol. 203, p. 104412). https://doi.org/10.1016/j.cognition.2020.104412

Brueckner, A. (2002). Williamson on the Primeness of Knowing. *Analysis*, *62*(3), 197–202. http://www.jstor.org/stable/3329200

Buckwalter, W., Rose, D., & Turri, J. (2015). Belief through Thick and Thin. *Noûs*, *49*(4), 748–775. https://doi.org/10.1111/nous.12048

Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition*, *112*(2), 337–342. https://doi.org/10.1016/j.cognition.2009.05.006

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language*, *28*(5), 606–637. https://doi.org/10.1111/mila.12036

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, *12*(5), 187–192. https://doi.org/10.1016/j.tics.2008.02.010

Carruthers, P., & Smith, P. K. (1996). *Theories of Theories of Mind*. Cambridge University Press. https://market.android.com/details?id=book-CtJ6BFChp9IC

Chakroff, A., Dungan, J., Koster-Hale, J., Brown, A., Saxe, R., & Young, L. (2016). When minds matter for moral judgment: intent information is neurally encoded for harmful but not impure acts. *Social Cognitive and Affective Neuroscience*, *11*(3), 476–484. https://doi.org/10.1093/scan/nsv131

Chang, W. (2009). Connecting counterfactual and physical causation. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 19983–11987. https://escholarship.org/content/qt61b6s1j1/qt61b6s1j1.pdf

Chisholm, R. (1966). *Theory of Knowledge. Englewood Cliffs, NJs Prentice-Hall*. Inc.

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature*, *386*(6625), 604–608. https://doi.org/10.1038/386604a0

Cushman, F., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition*, *108*(1), 281–289. https://doi.org/10.1016/j.cognition.2008.02.005

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science*, *35*(6), 1052–1075. https://doi.org/10.1111/j.1551-6709.2010.01167.x

Danks, D. (2017). Singular causation. *The Oxford Handbook of Causal Reasoning*, 201–215. https://books.google.com/books?hl=en&lr=&id=2qt0DgAAQBAJ&oi=fnd&pg=PA201&dq=Singular+causation&ots=azoq7o9M_U&sig=wkfIb0qHXLSOE_JmWx-xmI8u9gU

de Leeuw, J. R. (2015). jsPsych: a JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Dennett, D. C. (1978). Beliefs About Beliefs. *The Behavioral and Brain Sciences*, *1*(4), 568. https://philpapers.org/rec/DENBAB

Dennett, D. C. (1983). Taking the intentional stance seriously. *The Behavioral and Brain Sciences*, *6*(3), 379–390. https://doi.org/10.1017/s0140525x00016666

Dennis, S. A., Goodson, B. M., & Pearson, C. A. (2019). Online Worker Fraud and Evolving Threats to the Integrity of MTurk Data: A Discussion of Virtual Private Servers and the Limitations of IP-Based Screening Procedures. *Behavioral Research in Accounting*, *32*(1), 119–134. https://doi.org/10.2308/bria-18-044

Deschrijver, E., & Palmer, C. (2020). Reframing social cognition: Relational versus representational mentalizing. *Psychological Bulletin*, *146*(11), 941–969. https://doi.org/10.1037/bul0000302

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. In *NeuroImage* (Vol. 55, Issue 2, pp. 705–712). https://doi.org/10.1016/j.neuroimage.2010.12.040

Dowe, P. (2000). Causality and explanation. *The British Journal for the Philosophy of Science*, *51*(1), 165–174. https://doi.org/10.1093/bjps/51.1.165

Dowe, Phil. (1992). Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory. *Philosophy of Science*, *59*(2), 195–216. https://doi.org/10.1086/289662

Drayton, L. A., & Santos, L. R. (2018). What do monkeys know about others' knowledge? *Cognition*, *170*, 201–208. https://doi.org/10.1016/j.cognition.2017.10.004

Driver, J. (2008). Attributions of causation and moral responsibility. *Moral Psychology, Vol 2: The Cognitive Science of Morality: Intuition and Diversity.*, *2*(585), 423–439. https://psycnet.apa.org/fulltext/2007-14533-029.pdf

Dudley, R. (2018). Young children's conceptions of knowledge. *Philosophy Compass*, *13*(6), e12494. https://doi.org/10.1111/phc3.12494

Dudley, R., Orita, N., Hacquard, V., & Lidz, J. (2015). Three-year-Olds' understanding of know and think. In *Studies in Theoretical Psycholinguistics* (pp. 241–262). Springer International Publishing. https://doi.org/10.1007/978-3-319-07980-6_11

Dunbar, R. (2003). Psychology. Evolution of the social brain [Review of *Psychology. Evolution of the social brain*]. *Science*, *302*(5648), 1160–1161. https://doi.org/10.1126/science.1092116

Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, *6*(5), 178–190. https://onlinelibrary.wiley.com/doi/abs/10.1002/(SICI)1520-6505(1998)6:5%3C178::AID-EVAN5%3E3.0.CO;2-8

Dungan, J., & Saxe, R. (2012). Matched false-belief performance during verbal and nonverbal interference. *Cognitive Science*, *36*(6), 1148–1156. https://doi.org/10.1111/j.1551-6709.2012.01248.x

Fair, D. (1979). Causation and the Flow of Energy. *Erkenntnis. An International Journal of Analytic Philosophy*, *14*(3), 219–250. https://doi.org/10.1007/BF00174894

Fincham, F. D., & Jaspers, J. M. (1980). *Attribution of responsibility. From man the scientist to man the lawyer Advances in experimental social psychology Berkowitz, L*. Academic Press New York.

Fizke, E., Barthel, D., Peters, T., & Rakoczy, H. (2014). Executive function plays a role in coordinating different perspectives, particularly when one's own perspective is involved. *Cognition*, *130*(3), 315–334. https://doi.org/10.1016/j.cognition.2013.11.017

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, *7*(1), 1–9. https://doi.org/10.1057/s41599-020-0494-4

Flombaum, J. I., & Santos, L. R. (2005). Rhesus monkeys attribute perceptions to others. *Current Biology: CB*, *15*(5), 447–452. https://doi.org/10.1016/j.cub.2004.12.076

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology*, *63*, 287–313. https://doi.org/10.1146/annurev-psych-120710-100449

Frith, U. (2001). Mind blindness and the brain in autism. *Neuron*, *32*(6), 969–979. https://doi.org/10.1016/s0896-6273(01)00552-9

Gazdar, G. (1979). *Pragmatics: Implicature, Presupposition and Logical Form*. Academic Press. https://play.google.com/store/books/details?id=ourdtlC9K_4C

German, T. P., & Hehman, J. A. (2006). Representational and executive selection resources in "theory of mind": evidence from compromised belief-desire reasoning in old age. *Cognition*, *101*(1), 129–152. https://doi.org/10.1016/j.cognition.2005.05.007

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. https://doi.org/10.1037/rev0000281

Gerstenberg, T., Halpern, J. Y., & Tenenbaum, J. B. (2015). Responsibility judgments in voting scenarios. *CogSci*. http://web.mit.edu/tger/www/papers/Responsibility%20judgments%20in%20voting%20scenarios,%20Gerstenberg,%20Halpern,%20Tenenbaum,%202015.pdf

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-Tracking Causality. *Psychological Science*, *28*(12), 1731–1744. https://doi.org/10.1177/0956797617713053

Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*, *216*, 104842. https://doi.org/10.1016/j.cognition.2021.104842

Gettier, E. L. (1963). Is Justified True Belief Knowledge? In *Analysis* (Vol. 23, Issue 6, pp. 121–123). https://doi.org/10.1093/analys/23.6.121

Gobbini, M. I., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two takes on the social brain: a comparison of theory of mind tasks. *Journal of Cognitive Neuroscience*, *19*(11), 1803–1814. https://doi.org/10.1162/jocn.2007.19.11.1803

González, B., & Chang, L. J. (2021). Computational models of mentalizing. In K. Ochsner & M. Gilead (Eds.), *"The Neural Bases of Mentalizing"* (Vol. 1). Springer Press. https://doi.org/10.31234/osf.io/4tyd9

Grice, H. P. (1969). Utterer's Meaning and Intentions. *The Philosophical Review*, *78*(2), 147–177. https://www.pdcnet.org/phr/content/phr_1969_0078_0002_0147_0177

Hall, N. (2004). Two concepts of causation. *Causation and Counterfactuals*, 225–276. https://books.google.com/books?hl=en&lr=&id=JJL3xRNrCpUC&oi=fnd&pg=PA225&dq=Two+concepts+of+causation&ots=_fYyD2MjTj&sig=Cg3CgKe-ybMTGh_M0fpfU91bMtM

Hall, N. (2007). Causation. In *Oxford Handbooks Online*. https://doi.org/10.1093/oxfordhb/9780199234769.003.0019

Halpern, J. Y., & Hitchcock, C. (2013). Graded causation and defaults. In *arXiv [cs.AI]*. arXiv. https://doi.org/10.1093/bjps/axt050

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, *59*(4), 771–785. https://doi.org/10.1006/anbe.1999.1377

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour*, *61*(1), 139–151. https://doi.org/10.1006/anbe.2000.1518

Harris, P. L., Yang, B., & Cui, Y. (2017). 'I don't know': Children's early talk about knowledge. *Mind & Language*, *32*(3), 283–307. https://doi.org/10.1111/mila.12143

Hart, H. L. A., & Honoré, T. (1985). *Causation in the Law*. OUP Oxford. https://play.google.com/store/books/details?id=d7ZGAgAAQBAJ

Hausman, D. M. (2005). Causal Relata: Tokens, Types, or Variables? *Erkenntnis. An International Journal of Analytic Philosophy*, *63*(1), 33–54. https://doi.org/10.1007/s10670-005-0562-6

Hayashi, T., Akikawa, R., Kawasaki, K., Egawa, J., Minamimoto, T., Kobayashi, K., Kato, S., Hori, Y., Nagai, Y., Iijima, A., Someya, T., & Hasegawa, I. (2020). Macaques Exhibit Implicit Gaze Bias Anticipating Others' False-Belief-Driven Actions via Medial Prefrontal Cortex. *Cell Reports*, *30*(13), 4433-4444.e5. https://doi.org/10.1016/j.celrep.2020.03.013

Heim, I. (1991). Artikel und Definitheit. In *Semantik* (pp. 487–535). https://doi.org/10.1515/9783110126969.7.487

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. https://doi.org/10.1016/j.cognition.2019.05.006

Henne, P., Pinillos, Á., & De Brigard, F. (2017). Cause by Omission and Norm: Not Watering Plants. *Australasian Journal of Philosophy*, *95*(2), 270–283. https://doi.org/10.1080/00048402.2016.1182567

Hesslow, G. (1988). The problem of causal selection. *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*, 11–32. https://www.hesslow.com/GHNew/philosophy/Problemselection.htm

Hiddleston, E. (2005). A Causal Theory of Counterfactuals. *Noûs*, *39*(4), 632–657. http://www.jstor.org/stable/3506114

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. https://doi.org/10.1080/13546783.2016.1191547

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, *93*(1), 75–88. https://doi.org/10.1037/0033-295X.93.1.75

Hirschberg, J. L. B. (1985). *A theory of scalar implicature*. University of Pennsylvania.

Hitchcock, C., & Knobe, J. (2009). Cause and Norm. *The Journal of Philosophy*, *106*(11), 587–612. https://doi.org/10.5840/jphil20091061128

Hobson, R. P. (1984). Early childhood autism and the question of egocentrism. *Journal of Autism and Developmental Disorders*, *14*(1), 85–104. https://doi.org/10.1007/BF02408558

Hochstein, L., Bale, A., & Barner, D. (2018). Scalar Implicature in Absence of Epistemic Reasoning? The Case of Autism Spectrum Disorder. *Language Learning and Development: The Official Journal*

of the Society for Language Development, *14*(3), 224–240. https://doi.org/10.1080/15475441.2017.1343670

Holland, C., & Phillips, J. (2020). *A theoretically driven meta-analysis of implicit theory of mind studies: The role of factivity*. https://cogsci.mindmodeling.org/2020/papers/0387/0387.pdf

Horn, Laurence R. (1989). *A Natural History of Negation*. University of Chicago Press. https://philpapers.org/rec/HORANH

Horn, Laurence Robert. (1972). *ON THE SEMANTIC PROPERTIES OF LOGICAL OPERATORS IN ENGLISH* [University of California]. https://search.proquest.com/openview/817b0f83e5087931c11ac181857d1481/1?pq-origsite=gscholar&cbl=18750&diss=y&casa_token=eRhQ1GEOTPEAAAAA:WMvA4901bod4uWX3k3s8QnfoNF0ZBu6VZ9TxTCbw3qtn9f-QjP-__EYDt9fwXVerXzFvhXQBIN0

Horschler, D. J., MacLean, E. L., & Santos, L. R. (2020). Do Non-Human Primates Really Represent Others' Beliefs? In *Trends in Cognitive Sciences* (Vol. 24, Issue 8, pp. 594–605). https://doi.org/10.1016/j.tics.2020.05.009

Horschler, D. J., Santos, L. R., & MacLean, E. L. (2019). Do non-human primates really represent others' ignorance? A test of the awareness relations hypothesis. In *Cognition* (Vol. 190, pp. 72–80). https://doi.org/10.1016/j.cognition.2019.04.012

Hume, D. (1748). *An Enquiry Concerning Human Understanding* (R. by LaSalle, Ed.). Open Court Press.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93. https://doi.org/10.1016/j.cognition.2017.01.010

Ichikawa, J. J., & Steup, M. (2016). The analysis of knowledge. In *The Stanford encyclopedia of philosophy*. The Stanford Encyclopedia of Philosophy, fall.

James, W. (n.d.). *The Principles of Psychology*. https://doi.org/10.4324/9781912282494/analysis-william-james-principles-psychology-macat-team

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110. https://doi.org/10.1016/j.cobeha.2019.04.010

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604. https://doi.org/10.1016/j.tics.2016.05.011

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153. https://doi.org/10.1037/0033-295X.93.2.136

Kahneman, D., Slovic, S. P., Slovic, P., Tversky, A., & Cambridge University Press. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press. https://play.google.com/store/books/details?id=_0H8gwj4a1MC

Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but not what they believe. *Cognition*, *109*(2), 224–234. https://doi.org/10.1016/j.cognition.2008.08.010

Kano, F., Krupenye, C., Hirata, S., Tomonaga, M., & Call, J. (2019). Great apes use self-experience to anticipate an agent's action in a false-belief test. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(42), 20904–20909. https://doi.org/10.1073/pnas.1910095116

Karg, K., Schmelz, M., Call, J., & Tomasello, M. (2015). The goggles experiment: can chimpanzees use self-experience to infer what a competitor can see? *Animal Behaviour*, *105*, 211–221. https://doi.org/10.1016/j.anbehav.2015.04.028

Karttunen, L. (1977). Syntax and semantics of questions. *Linguistics and Philosophy*, *1*(1), 3–44. https://doi.org/10.1007/BF00351935

Kiparsky, P., & Kiparsky, C. (2014). FACT. In *Progress in Linguistics* (pp. 143–173). De Gruyter Mouton. https://www.degruyter.com/document/doi/10.1515/9783111350219.143/html?lang=de

Kirfel, L., & Lagnado, D. (2021). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition*, *212*, 104721. https://doi.org/10.1016/j.cognition.2021.104721

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

Kissine, M. (2012). Pragmatics, cognitive flexibility and autism spectrum disorders. *Mind & Language*, *27*(1), 1–28. https://doi.org/10.1111/j.1468-0017.2011.01433.x

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. https://doi.org/10.1016/j.shpsa.2009.03.009

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. https://doi.org/10.1111/cogs.12792

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209. https://doi.org/10.1016/j.cognition.2015.01.013

Kominsky, J. F., & Scholl, B. J. (2020). Retinotopic adaptation reveals distinct categories of causal perception. *Cognition*, *203*(104339), 104339. https://doi.org/10.1016/j.cognition.2020.104339

Koster-Hale, J., Bedny, M., & Saxe, R. (2014). Thinking about seeing: perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. *Cognition*, *133*(1), 65–78. https://doi.org/10.1016/j.cognition.2014.04.006

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: a neural prediction problem. *Neuron*, *79*(5), 836–848. https://doi.org/10.1016/j.neuron.2013.08.020

Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: susceptibility to others' beliefs in human infants and adults. *Science*, *330*(6012), 1830–1834. https://doi.org/10.1126/science.1190792

Krachun, C., Carpenter, M., Call, J., & Tomasello, M. (2009). A competitive nonverbal false belief task for children and apes. *Developmental Science*, *12*(4), 521–535. https://doi.org/10.1111/j.1467-7687.2008.00793.x

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114. https://doi.org/10.1126/science.aaf8110

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive Physics: Current Research and Controversies. *Trends in Cognitive Sciences*, *21*(10), 749–759. https://doi.org/10.1016/j.tics.2017.06.002

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: the effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770. https://doi.org/10.1016/j.cognition.2008.06.009

Lakoff, R., Bierwisch, M., & Heidolph, K. E. (1973). Progress in Linguistics: A Collection of Papers. In *Language* (Vol. 49, Issue 3, p. 685). https://doi.org/10.2307/412359

Lakoff, R. T. (1968). *Abstract syntax and latin complementation* (Vol. 26, pp. 383–421). MIT Press. https://doi.org/10.1016/0024-3841(71)90004-0

Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., & Gerstenberg, T. (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology*, *129*, 101412. https://doi.org/10.1016/j.cogpsych.2021.101412

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

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in 'theory of mind.' *Trends in Cognitive Sciences*, *8*(12), 528–533. https://www.sciencedirect.com/science/article/pii/S1364661304002608

Leslie, A. M., Knobe, J., & Cohen, A. (2006). Acting Intentionally and the Side-Effect Effect: Theory of Mind and Moral Judgment. *Psychological Science*, *17*(5), 421–427. https://doi.org/10.1111/j.1467-9280.2006.01722.x

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*. https://www.jstor.org/stable/2215339

Lewis, David. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, *50*(3), 249–258. https://doi.org/10.1080/00048407212341301

Lewis, David. (1973). Counterfactuals and Comparative Possibility. In *IFS* (pp. 57–85). Springer Netherlands. https://doi.org/10.1007/978-94-009-9117-0_3

Lewis, David. (1974). Causation. *The Journal of Philosophy*, *70*(17), 556–567. https://doi.org/10.2307/2025310

Lipe, M. G. (1991). Counterfactual reasoning as a framework for attribution theories. *Psychological Bulletin*, *109*(3), 456–471. https://doi.org/10.1037/0033-2909.109.3.456

Liu, S., & Spelke, E. S. (2017). Six-month-old infants expect agents to minimize the cost of their actions. *Cognition*, *160*, 35–42. https://doi.org/10.1016/j.cognition.2016.12.007

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041. https://doi.org/10.1126/science.aag2132

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332. https://doi.org/10.1016/j.cogpsych.2010.05.002

Luo, Y., & Johnson, S. C. (2009). Recognizing the role of perception in action at 6 months. *Developmental Science*, *12*(1), 142–149. https://doi.org/10.1111/j.1467-7687.2008.00741.x

Lyons, J. (2017). Epistemological Problems of Perception. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2017/entries/perception-episprob/

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, *67*(1), 1–25. https://doi.org/10.1086/392759

Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., Sirker, S., Usui, N., & Hashimoto, T. (2017). Gettier across cultures 1. *Nous* , *51*(3), 645–664. https://doi.org/10.1111/nous.12110

Mackie, J. L. (1980). *The Cement of the Universe: A Study of Causation*. Clarendon Press. https://play.google.com/store/books/details?id=4QdREAAAQBAJ

Malle, B. F. (2021). Moral Judgments. In *Annual Review of Psychology* (Vol. 72, Issue 1, pp. 293–318). https://doi.org/10.1146/annurev-psych-072220-104358

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A Theory of Blame. *Psychological Inquiry*, *25*(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

Martin, A. (2019). Belief Representation in Great Apes [Review of *Belief Representation in Great Apes*]. *Trends in Cognitive Sciences*, *23*(12), 985–986. https://doi.org/10.1016/j.tics.2019.10.008

Martin, A., & Santos, L. R. (2014). The origins of belief representation: monkeys fail to automatically represent others' beliefs. *Cognition*, *130*(3), 300–308. https://doi.org/10.1016/j.cognition.2013.11.016

Martin, A., & Santos, L. R. (2016). What Cognitive Representations Support Primate Theory of Mind? *Trends in Cognitive Sciences*, *20*(5), 375–382. https://doi.org/10.1016/j.tics.2016.03.005

McCloskey, M. (1983). Intuitive Physics. *Scientific American*, *248*(4), 122–131. http://www.jstor.org/stable/24968881

McClure, J., Hilton, D. J., & Sutton, R. M. (2007). Judgments of voluntary and physical causes in causal chains: probabilistic and social functionalist criteria for attributions. *European Journal of Social Psychology*, *37*(5), 879–901. https://doi.org/10.1002/ejsp.394

McGrath, S. (2005). Causation by Omission: A Dilemma. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *123*(1/2), 125–148. http://www.jstor.org/stable/4321576

McKinnon, M. C., & Moscovitch, M. (2007). Domain-general contributions to social reasoning: theory of mind and deontic reasoning re-explored. *Cognition*, *102*(2), 179–218. https://doi.org/10.1016/j.cognition.2005.12.011

Menzies, P. (2004). Causal Models, Token Causation, and Processes. *Philosophy of Science*, *71*(5), 820–832. https://doi.org/10.1086/425057

Meyer, M. L., Taylor, S. E., & Lieberman, M. D. (2015). Social working memory and its distinctive link to social cognitive ability: an fMRI study. *Social Cognitive and Affective Neuroscience*, *10*(10), 1338–1347. https://academic.oup.com/scan/article-abstract/10/10/1338/1650896

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

Moore, C., Bryant, D., & Furrow, D. (1989). Mental Terms and the Development of Certainty. *Child Development*, *60*(1), 167–171. https://doi.org/10.2307/1131082

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. D. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(7), 2688–2692. https://doi.org/10.1073/pnas.1011734108

Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PloS One*, *14*(8), e0219704. https://doi.org/10.1371/journal.pone.0219704

Myers-Schulz, B., & Schwitzgebel, E. (2013). Knowing that P without believing that P. *Noûs*. https://onlinelibrary.wiley.com/doi/abs/10.1111/nous.12022

Nagel, J. (2017). Factive and nonfactive mental state attribution. *Mind & Language*, *32*(5), 525–544. https://doi.org/10.1111/mila.12157

Newton, A. M., & de Villiers, J. G. (2007). Thinking while talking: adults fail nonverbal false-belief reasoning. *Psychological Science*, *18*(7), 574–579. https://doi.org/10.1111/j.1467-9280.2007.01942.x

Oktay-Gür, N., & Rakoczy, H. (2017). Children's difficulty with true belief tasks: Competence deficit or performance problem? *Cognition*, *166*, 28–41. https://doi.org/10.1016/j.cognition.2017.05.002

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258. https://doi.org/10.1126/science.1107621

Paul, L. A. (1998). Problems with Late Preemption. *Analysis*, *58*(1), 48–53. http://www.jstor.org/stable/3328155

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press. https://play.google.com/store/books/details?id=wnGU_TsW3BQC

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. https://doi.org/10.1037/a0021523

Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., Santos, L., & Knobe, J. (2020). Knowledge before belief. In *Behavioral and Brain Sciences* (Vol. 44). https://doi.org/10.1017/s0140525x20000618

Phillips, J., & George, B. R. (2018). Knowledge wh and False Beliefs: Experimental Investigations. *Journal of Semantics*, *35*(3), 467–494. https://doi.org/10.1093/semant/ffy004

Phillips, J., & Norby, A. (2019). Factive theory of mind. *Mind & Language*, *36*(1), 3–26. https://doi.org/10.1111/mila.12267

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/s0140525x00076512

Quillien, T. (2020). When do we think that X caused Y? *Cognition*, *205*(104410), 104410. https://doi.org/10.1016/j.cognition.2020.104410

Quillien, T., & German, T. C. (2021). A simple definition of "intentionally." *Cognition*, *214*, 104806. https://doi.org/10.1016/j.cognition.2021.104806

Rolfs, M., Dambacher, M., & Cavanagh, P. (2013). Visual adaptation of the perception of causality. *Current Biology: CB*, *23*(3), 250–254. https://doi.org/10.1016/j.cub.2012.12.017

Rose, D., & Schaffer, J. (2013). Knowledge entails dispositional belief. *Philosophical Studies*, *166*(1), 19–50. https://doi.org/10.1007/s11098-012-0052-z

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press. https://play.google.com/store/books/details?id=2ug9DwAAQBAJ

Salmon, W. C. (1994). Causality without counterfactuals. *Philosophy of Science*, *61*(2), 297–312. https://doi.org/10.1086/289801

Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition*, *156*, 164–176. https://doi.org/10.1016/j.cognition.2016.07.007

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*. https://doi.org/10.1023/B:LING.0000023378.71748.db

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, *55*(1), 87–124. https://doi.org/10.1146/annurev.psych.55.090902.142044

Schaffer, J. (2005). Contrastive Causation. *The Philosophical Review*, *114*(3), 327–358. http://www.jstor.org/stable/30043679

Schultz, J., Friston, K. J., O'Doherty, J., Wolpert, D. M., & Frith, C. D. (2005). Activation in posterior superior temporal sulcus parallels parameter inducing the percept of animacy. *Neuron*, *45*(4), 625–635. https://doi.org/10.1016/j.neuron.2004.12.052

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome. *Science*, *325*(5942), 883–885. https://doi.org/10.1126/science.1176170

Sosa, F. A., Ullman, T., Tenenbaum, J. B., Gershman, S. J., & Gerstenberg, T. (2021). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*, *217*, 104890. https://doi.org/10.1016/j.cognition.2021.104890

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, *27*(1), 76–105. https://doi.org/10.1016/0022-1031(91)90011-T

Starmans, C., & Friedman, O. (2012). The folk conception of knowledge. *Cognition*, *124*(3), 272–283. https://doi.org/10.1016/j.cognition.2012.05.017

Stich, S. (2013). Do Different Groups Have Different Epistemic Intuitions? A Reply to Jennifer Nagel1. *Philosophy and Phenomenological Research*, *87*(1), 151–178. https://doi.org/10.1111/j.1933-1592.2012.00590.x

Suppes, P. (1968). *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Pub. Co. https://philpapers.org/rec/SUPAPT

Sytsma, J., Livengood, J., & Rose, D. (2012). Two types of typicality: rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(4), 814–820. https://doi.org/10.1016/j.shpsc.2012.05.009

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. https://doi.org/10.1016/j.tics.2017.05.012

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive Psychology*, *104*, 57–82. https://doi.org/10.1016/j.cogpsych.2017.05.006

Vouloumanos, A., Martin, A., & Onishi, K. H. (2014). Do 6-month-olds understand that speech can communicate? *Developmental Science*, *17*(6), 872–879. https://doi.org/10.1111/desc.12170

Walsh, C. R., & Sloman, S. A. (2011). The meaning of cause and prevent: The role of causal mechanism. *Mind & Language*, *26*(1), 21–52. https://doi.org/10.1111/j.1468-0017.2010.01409.x

Waskan, J. (2011). Mechanistic explanation at the limit. *Synthese*, *183*(3), 389–408. https://doi.org/10.1007/s11229-010-9869-1

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief. In *Child Development* (Vol. 72, Issue 3, pp. 655–684). https://doi.org/10.1111/1467-8624.00304

Westra, E., & Nagel, J. (2021). Mindreading in conversation. *Cognition*, *210*, 104618. https://doi.org/10.1016/j.cognition.2021.104618

Williamson, T. (2002). *Knowledge and Its Limits*. Oxford University Press. https://play.google.com/store/books/details?id=tMDqMUTg6gYC

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128. https://www.ncbi.nlm.nih.gov/pubmed/6681741

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology. General*, *136*(1), 82–111. https://doi.org/10.1037/0096-3445.136.1.82

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. https://doi.org/10.1037/a0018129

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, USA. https://play.google.com/store/books/details?id=IFVJAbgySmEC

Yeung, S. K., Yay, T., & Feldman, G. (2022). Action and Inaction in Moral Judgments and Decisions: Meta-Analysis of Omission Bias Omission-Commission Asymmetries. *Personality & Social Psychology Bulletin*, *48*(10), 1499–1515. https://doi.org/10.1177/01461672211042315

Yuan, Y., & Kim, M. (2021). Cross-Cultural Convergence of Knowledge Attribution in East Asia and the US. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-021-00523-y

Zaitchik, D., Walker, C., Miller, S., LaViolette, P., Feczko, E., & Dickerson, B. C. (2010). Mental state attribution and the temporoparietal junction: an fMRI study comparing belief, emotion, and perception. *Neuropsychologia*, *48*(9), 2528–2536. https://doi.org/10.1016/j.neuropsychologia.2010.04.031

Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: causality and counterfactuals in group attributions. *Cognition*, *125*(3), 429–440. https://doi.org/10.1016/j.cognition.2012.07.014