A simple definition of "intentionally"

Tadeg Quillien & Tamsin German

Abstract: Cognitive scientists have been debating how the folk concept of intentional action works. We suggest a simple account: people consider that an agent did X *intentionally* to the extent that X was causally dependent on how much the agent wanted X to happen (or not to happen). Combined with recent models of human causal cognition, this definition provides a good account of the way people use the concept of intentional action, and offers natural explanations for puzzling phenomena such as the side-effect effect. We provide empirical support for our theory, in studies where we show that people's causation and intentionality judgments track each other closely, in everyday situations as well as in scenarios with unusual causal structures. Study 5 additionally shows that the effect of norm violations on intentionality judgments depends on the causal structure of the situation, in a way uniquely predicted by our theory. Taken together, these results suggest that the folk concept of intentional action has been difficult to define because it is made of cognitive building blocks, such as our intuitive concept of causation, whose logic cognitive scientists are just starting to understand.

Keywords: intentional action; causation; concepts; theory of mind

1) Introduction

Much of cognition works under the radar of consciousness. This puts us in a strange position: even though we know the meaning of the words we use, often we are unable to give them explicit definitions. This has caused many debates over the precise meaning of certain concepts. These debates can be interesting because the way we use words provides important clues about the hidden format of mental representations (Pinker, 2007; Strickland, 2017).

As an example, what do we mean when we say that someone did something "intentionally"? We all have an intuitive understanding of the concept, yet a lot of ink has been spilled by cognitive scientists searching for an explicit definition (e.g., Bennett, 1965; Searle, 1983; Davidson, 1980; Bratman, 1984; Mele, 2001; Malle & Knobe, 1997; Cushman & Mele, 2008; Cova, Dupoux & Jacob, 2012). So far, no strong consensus has been reached – why?

Maybe the algorithms that our brain uses in order to decide whether something is intentional are very complicated, or maybe there are not one but several different concepts of intentional action (e.g. Malle & Knobe, 1997; Cushman & Mele, 2008; Cova, Dupoux & Jacob, 2012). Here we defend an alternative approach: the folk concept of intentional action is relatively simple, but looks complicated because it is made out of building blocks the structure of which is not completely obvious *a priori*.

We suggest that the concept of intentional action is at its core, a *causal* concept. Roughly, an agent did X *intentionally* to the extent that X was causally dependent on how much the agent wanted X to happen (or not to happen). This hypothesis stems from the fact that causation is at the core of our commonsense psychology; the way we explain and predict the behavior of others relies on a mental causal model of how mental states and states of the world interact with one

another. Therefore, it makes sense that causation would be central to folk concepts about the mind.

We want to show that many features of the concept of intentional action emerge naturally from our simple theory, given i) the way that commonsense psychology works, and ii) the way that the human mind represents causation. In other words, we are trying to understand the concept by making a hypothesis about its basic building blocks, and looking at what cognitive scientists know about these building blocks. This means that our theory can only be as good as our current scientific understanding of the building blocks involved. Nonetheless, we hope to show that it can illuminate several puzzling phenomena, and offers new fruitful predictions about the way people use the word "intentionally".

We proceed as follows. First, we briefly review existing accounts of intentional action. Second, we highlight relevant research on commonsense psychology and causal cognition, and from that research derive our definition of intentional action. Third, we show how this simple account can explain many known features of people's use of "intentionally". Fourth, we present the results of six studies testing predictions of the account. In study 1, we show that our definition closely tracks people's intuitions about intentionality in everyday cases. In study 2, we examine a case where traditional philosophical analyses of causation hold that the agent's desires caused the outcome, yet people do not think the agent acted intentionally (cases of so-called 'deviant causation'). We show that in such cases, people's causation judgments are actually almost as low as their intentionality judgments. In study 3, we show that people think that agents can act intentionally even when the agent has only a very weak belief that their action will lead to the outcome – this is consistent with our account, but inconsistent with standard theories which posit that belief is a central pre-requisite of intentionality. In study 4, we show that causal

judgments exhibit a "side-effect effect" which parallels that observed for intentionality. Study 5 shows that statistical norms interact with the causal structure of a situation to shape intentionality judgments, in a way uniquely predicted by recent models of causal cognition. Finally, study 6 demonstrates that in a case where our account predicts a dissociation between intentionality and causation judgments, they do indeed come apart.

2) Existing accounts of intentional action

In the Philosophical Investigations, Wittgenstein (1953) asks, "What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?". The puzzle of what makes an event an intentional action has attracted a large amount of attention from philosophers (e.g., Anscombe, 1957; Davidson, 1980; Mele, 2009).

In parallel, psychologists have been interested in the concept because perceptions of intentionality play an important role in social cognition. For example, whether we perceive an action as intentional influences how we form impressions about the actor (Heider, 1958), how we judge the action morally and legally (Cushman, 2015), and the nature and intensity of the emotions invoked by the action (Sell et al., 2017; Tesser et al., 1968).

Intuitively, it seems easy to generate a list of criteria that an action must exhibit in order to count as intentional. But these list-based definitions are vulnerable to counter-examples. It is then tempting to deal with these counter-examples by adding new criteria to the original list. As a result, there has been a steady increase in the complexity of theories of intentional action over the years (as documented by Malle & Knobe, 1997).

Early philosophical accounts (Aristotle, 330BC; Hume, 1740) put forward two criteria for acting intentionally: one needs to have a *desire* for the outcome, and a *belief* that the act would

lead to the outcome. The two-components theory was later found lacking: one can imagine (for example) a basketball player who wants to win the game, and thinks that fouling would help her achieve that goal, yet does not foul intentionally when she does.

Accordingly, later theories were three-component models: they stipulated that beliefs and desires must jointly cause an *intention* to act (Brand, 1984; Bratman, 1987; Searle, 1983; Thalberg, 1984). In parallel, social psychologists identified a fourth component of intentionality: an agent needs some degree of *skill* (or control, ability) in carrying out the action (Heider, 1958; Shaver, 1985; Ossorio and Davis, 1968; Jones & Davis, 1965). For instance, a novice dart player who hits a difficult target due to pure luck did not *intentionally* hit the target, despite the fact that he wanted to (Malle & Knobe, 1997; Knobe 2003b).

To probe people's explicit concept of intentionality, Malle & Knobe (1997) asked undergraduate students to write down their definition of what it means for someone to do something intentionally. They found that the explicit folk concept of intentionality contains the four components identified above, as well as a fifth component, *awareness* of what one is doing while doing it.

To make matters worse, it was later discovered that moral considerations can have a profound influence on people's attributions of intentionality (Knobe, 2003a,b). For instance, people judge that a CEO who harms the environment as a side-effect of implementing a new policy did so intentionally, despite the fact that the CEO was indifferent toward that side-effect. By contrast, people do not judge that a CEO who *helps* the environment as a side-effect does so intentionally. This result is inconsistent with the five-component model, which predicts that in both cases the CEO should be viewed as not having intentionally caused the side-effect.

The discovery prompted an avalanche of research aimed at explaining this "side-effect effect" (e.g. Nadelhoffer, 2006; Wright & Bengson, 2009, Pettit & Knobe, 2009; Uttich & Lombrozo, 2010; Cova, Dupoux & Jacob, 2012; Adams & Steadman, 2004; Hindricks, 2014; Sripada, 2012; Sloman, Fernbach & Ewing, 2012; Machery, 2008; Leslie, Knobe & Cohen, 2006)[1]. On one account, people may consider that an agent does something intentionally when the agent's attitude towards an outcome exceeds a given threshold, and moral considerations influence where people put this threshold (Knobe, 2010). On another theory, the concept of intentional action might be fundamentally sensitive to whether people comply with the normative reasons for or against acting (Hindriks, 2014). On yet other accounts, the side-effect effect is not intrinsically about morality, because similar effects arise in non-moral scenarios. For instance, people judge that an agent who pays $1 extra to get an extra-large beverage *intentionally* pays the extra money (despite the fact that paying extra money was not the agent's goal in ordering the extra-large beverage; Machery, 2008). According to Machery (2008), people view side-effects as intentional when there is a trade-off between the costs generated by this side effect and the benefits of the primary goals of the action.

Yet, as Cova (2016) notes, most of the theories that aim to explain the side-effect effect (or related empirical findings) are relatively limited in scope: they usually account for, at best, a handful of empirical phenomena about how people use the concept of intentional action, but remain silent about, or are inconsistent with, other features of the concept.

[1] Note that some researchers argue that the effect actually tells us nothing about the folk concept of intentional action. Instead of reflecting people's core concept, it arises because motivated reasoning (Alicke & Rose, 2010), or the pragmatics of ordinary conversation (Adams & Steadman, 2004) makes people use "intentionally" as a way to imply blameworthiness. Or perhaps the effect shows that the emotions we feel when evaluating a situation distort our ability to correctly use the concept (Nadelhoffer, 2006). See Knobe (2010) for arguments against such interpretations.

The lack of prospect for a unified theory of intentional action has even led some researchers to suggest that there is none to be found. Instead, they argue, there might actually be several distinct concepts of intentionality, each of them invoked depending on the context at hand. For instance, Nichols & Ulatowski (2007) suggest that we sometimes use "S intentionally did X" to mean "S had a motive to do X", and sometimes to mean "S knew that his action would result in X" (for other polysemic theories, see Sousa & Holbrook, 2010; Cushman & Mele, 2008; Cova, Dupoux & Jacob, 2012).

We think that these difficulties may be explained by the fact that current approaches tend to follow an inductive strategy. That is, researchers start from the intuitions that people have about intentional action, and try to construct an account that fits these intuitions. Here we take a theory-driven approach instead. We start from what cognitive scientists know about the mechanisms via which people reason about the mind, and we ask: "if a concept of intentional action emerged from the operation of these mechanisms, what would it look like?".

Our theory shares similarities with *causalist* approaches in the philosophy of action (Mele, 2009), notably that of Donald Davidson (1980). According to Davidson, what makes an event an intentional action is the fact that it was jointly *caused* by the relevant beliefs and desires of the agent.

Causalist approaches have traditionally had difficulty dealing with cases of 'causal deviance': scenarios where an agent's beliefs and desires jointly cause an outcome but that few people would consider as involving intentional action (see section 4.1 for examples). The existence of such cases led Davidson to specify that causation must happen "in the right way" to count as intentional; but he did not provide a theory of what makes a causal link the right kind of causal link. Indeed, he considered that such a theory would be a matter of empirical discovery:

facts about the way the mind works ultimately determine what it means for mental causation to count as intentional (Davidson, 1980; see Goldman, 1970, for a similar view). Unlike Davidson, here we are not trying to give a philosophical or scientific definition of intentional action: we are interested in the folk concept. This means that it is easier for us to actually provide a theory of what counts as causation "in the right way", since we only need to determine what counts as such according to commonsense psychology.

3) Building blocks

3.1) Commonsense psychology

As part of their mental toolkit, humans are equipped with a set of reliably-developing cognitive mechanisms that allow them to predict and explain the behavior of others – collectively, these are referred to as Theory of Mind, or commonsense psychology (Dennett, 1987; Leslie, 1994; Leslie, Friedman & German, 2004; Baillargeon, Scott & Bian, 2016). There are many competing theories of commonsense psychology, but most of them share the idea that it is essentially a causal inference engine: it leverages causal knowledge in order to generate inferences about people's mental states and their behavior (Dennett, 1987; Gopnik & Wellman, 1992; Leslie, 1994; Apperly & Butterfill, 2009; Baker, Jara-Ettinger, Saxe & Tenenbaum, 2017).

In other words, at a computational level of analysis (Marr, 1982), we can think of commonsense psychology as relying on an internal causal model of the way mental states and states of the world interact with each other. Commonsense psychology consists of a set of inference algorithms that leverage this causal model to make a variety of useful inferences (such as predicting an agent's behavior given its mental states, or vice-versa).

Here we are interested in the part of this causal model that is used to predict and explain an agent's behavior given its mental states. A popular idea has been that this part of commonsense psychology relies on the kind of causal model depicted in Diagram 1a: people have beliefs and desires, which jointly cause their actions (Davidson, 1963; Dennett, 1987; Wertz & German, 2007).

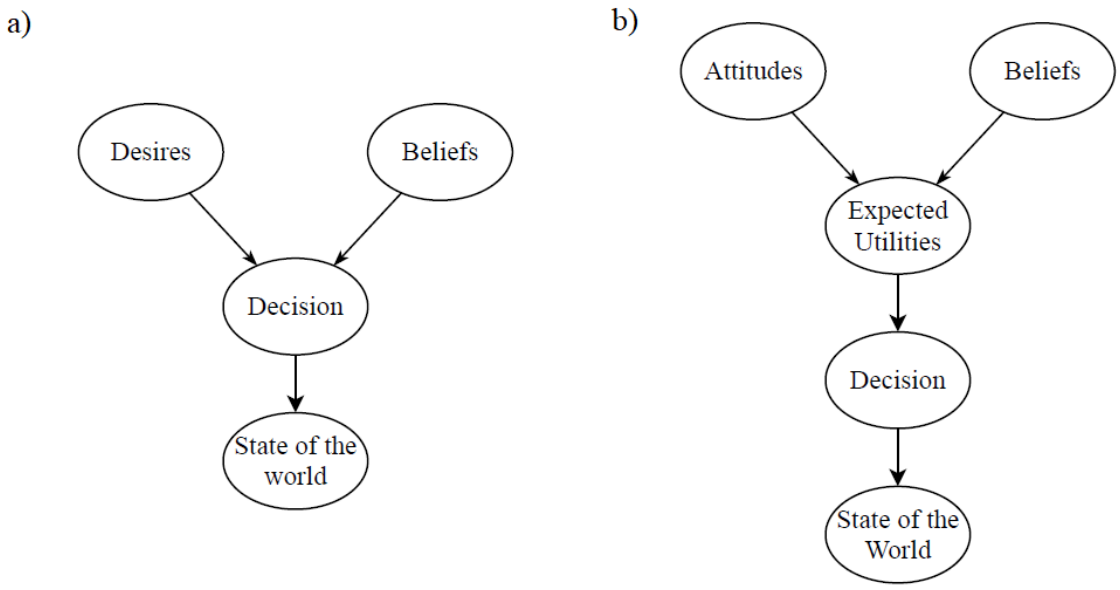

Diagram 1. a) The simple 'Belief+Desire' causal model. b) The generative causal model suggested by an expected utility framework. "Attitudes" determine how much the agent values a given state of the world; an attitude can be positive or negative.

In recent years, cognitive scientists have conducted extensive empirical and modeling work to refine our understanding of this causal model. Their work suggests that people explain the behavior of others in roughly the same way an economist would: from an early age, people

spontaneously model agents as expected-utility maximizers that behave in an approximately rational way given their beliefs and desires (Woodward, 1998; Gergely & Csibra, 2003; Baker, Saxe & Tenenbaum, 2009; Scott & Baillargeon, 2013; Johnson & Rips, 2015; Jara-Ettinger, Gweon, Schulz & Tenenbaum, 2016; Lucas et al., 2014; Liu, Ullman, Tenenbaum & Spelke, 2017; Jara-Ettinger, Schultz & Tenenbaum, 2020).

An expected utility framework suggests a computationally explicit causal model of how people make decisions (sketched in diagram 1b; see Jara-Ettinger, Schulz & Tenenbaum, 2020 for a more detailed computational model, and empirical tests of its fit to human intuitions)[2]. Just as the simple causal model shown in diagram 1a, it partitions the relevant mental states into a motivational and an epistemic component.

At the motivational level, agents assign different values (utilities) to different states of the world; for instance an agent may assign a positive value to getting food, and a negative value to receiving electric shocks. We will refer to these value assignments as *attitudes*: an agent has a positive attitude toward an outcome if it assigns a positive value to that outcome, and a negative attitude if it assigns a negative value to the outcome. Thus, to a rough approximation attitudes toward an outcome can be seen as lying on a continuum from extremely negative to extremely positive (see Pettit & Knobe, 2009). [3] Henceforth we will use this notion of attitude because it is

[2] By "expected utility framework", we have in mind something broader than expected utility theory *stricto sensu* (Von Neumann & Morgenstern, 1944; Savage, 1954). Expected utility theory assumes that agents obey strict standards of rationality (for instance, they have transitive preferences). There are other theories of decision-making that model agents as expected utility maximizers, without assuming that they obey every axiom of rationality (e.g., prospect theory, Kahneman & Tversky, 1979). The argument we make here does not depend on the exact extent to which people assume other agents to be rational.

[3] Although note that at a more mechanistic level, commonsense psychology might represent negative and positive attitudes differently at certain stages of processing; see Leslie & Polizzi, 1998; Leslie, German & Polizzi, 2005.

more computationally explicit than the naive concept of 'desire'. In particular, 'desire' tends to denote a positive attitude toward an outcome, whereas people are also able to represent the negative attitudes an agent may have toward an outcome.

According to an expected utility framework, beliefs and attitudes jointly determine an agent's decisions by determining the expected utility that the agent assigns to a given action. The expected utility of an action is a weighted sum of the utility of all possible outcomes of the action, where the utility of an outcome is weighted by its estimated probability. Then, the agent selects a course of action according to some procedure where actions with a higher expected utility are more likely to be selected.

This framework motivates our suggestion that, to the mind, intentionality is about the existence of a causal relationship between an agent's attitude toward a state of the world and that state of the world obtaining.

We can now formulate our hypothesis:

For the human mind, an agent did X intentionally if the agent's attitude toward X caused X, and caused X according to the typical causal model implicit in our commonsense psychology.

Here are two examples[4]:

Window. "Anne opens the window in order to let sunlight into the room. She believed that opening the window would let sunlight into the room, and had a positive attitude toward that

[4] For the sake of argument, here we assume that commonsense psychology does contain a causal model resembling the one sketched in diagram 1b. The general principle stated above could work with alternative versions of this causal model – the exact form this model takes is ultimately a matter of empirical discovery

outcome. This attitude – belief pair led her to compute a high expected utility for the action of opening the window, and she chose to do so as a result."

Anne's desire to let sunlight into the room caused sunlight to enter the room, and caused it in the typical way specified by commonsense psychology. Accordingly, it feels natural to say that Anne intentionally let sunlight into the room.

King. "The king's advisors have put a high-tech brain sensor on the king's head, which gives them a direct readout on what the king wants. Wondering whether they should build a bridge over the river, the advisors consult the brain sensor, and thereby learn that the king would be in favor of building a bridge if he were to be asked. Now that they know how the king feels, they go ahead and build the bridge, without bothering to formally ask him.

Here, the king's desire to build a bridge caused the bridge to be built, but this causal path clearly deviates from the way that, according to diagram 1b, desires cause outcomes: the king's desire did not cause the bridge to be built via a decision that the king made. In order to model the causal structure of this scenario, one needs to add extra causal links to the typical causal model depicted in Diagram 1b; see Diagram 2. So, intuitively the king did not intentionally build the bridge.

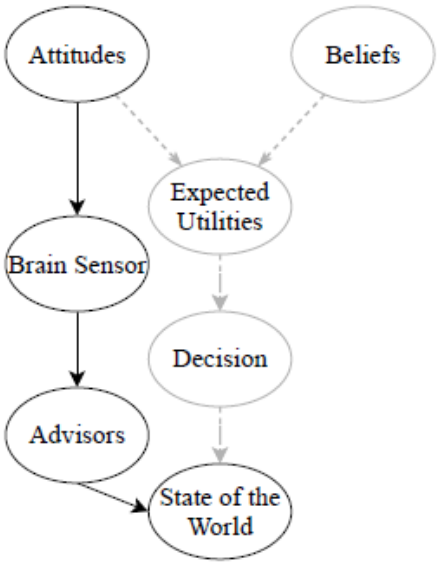

Diagram 2. Causal model for the **King** scenario. Attitudes affect the state of the world via a causal pathway (black solid arrows), which is different from the one pre-specified by commonsense psychology (grey dashed arrows).

3.2) Causal judgment

Most of our causal knowledge is embedded in domain-specific cognitive mechanisms (such as commonsense psychology), but we also clearly possess an abstract concept of causation that we can apply across domains: people spontaneously use words like “cause” and “because” to talk about almost anything (Gerstenberg & Tenenbaum, 2017; Icard et al., 2017; Quillien, 2020).

To the human mind, causation is a matter of counterfactual dependence. “C caused E” means that in a counterfactual alternative scenario where we ‘force’ C to not occur, E does not occur either (Gerstenberg, Peterson, Goodman, Lagnado & Tenenbaum, 2017; for philosophical analyses of causation along these lines see e.g. Lewis, 1973; Hitchcock, 2001; Halpern & Pearl,

2005; Weslake, 2015)[5] [6]. Consider a forest that catches on fire after a lightning bolt strikes a tree. If the lightning had not struck, the forest would not have caught fire: therefore the lightning bolt is a cause of the fire. Similarly, if there was no oxygen in the air to fuel the combustion, the fire would not have started: therefore oxygen is a cause of the fire.

Most philosophical analyses of causation have been *egalitarian*: they do not discriminate between different causes (Mill, 1856; Hall, 2004; Lewis, 1973), and would regard the lightning bolt and the oxygen as equally good causes of the fire. By contrast, our intuitive sense of causation does discriminate between causes (Hart & Honoré, 1985; Morris et al., 2019). Most people view the lightning bolt, rather than the oxygen, as the main cause of the fire. Similarly, most people think that a presidential candidate won the election because he won the swing state, not because he won the party stronghold (Quillien & Barlev, 2021).

Recently, cognitive scientists have made progress in understanding what drives gradation in causal judgment (Icard et al., 2017; Quillien, 2020; Gerstenberg, Lagnado, Goodman & Tenenbaum, 2021)[7].

When people think about the causes of an event, they implicitly consider several different counterfactuals to the event. For example, when thinking about the forest fire, people may consider other possible versions of the event. These can include a counterfactual where there is

[5] Contemporary models of causation, though they rely heavily on this counterfactual criterion, are of course more complicated. They are also designed to deliver the correct intuition in cases where a factor is not necessary for the effect. For example, when two soldiers in a firing squad fire at a prisoner at the same time neither soldier is individually necessary for the prisoner's death, although it makes sense to think of each soldier as a cause of the prisoner's death; see Halpern (2016) for review..

[6] In this paper, we focus on so-called 'dependence' theories of causation. Some philosophers (Schaffer, 2000; Dowe, 2004) and psychologists (Wolff, 2007) favor 'process' theories, which view causation as being (or being represented as) a physical exchange of entities between events – but we lack the space to discuss them.

[7] Although for some alternatives to the general framework described here, see Wolff, 2007; Alicke, Rose & Bloom, 2011; Sytsma, Livengood & Rose, 2012.

no lightning bolt (and the fire does not start), a counterfactual where the wind is stronger (and the fire spreads even faster), a counterfactual where the ground is wet, etc. People tend to think that C is a cause of E to the extent that C and E are highly *correlated* across these counterfactuals (Quillien, 2020; for empirical evidence consistent with this account see, e.g., Lombrozo, 2010; Icard et al., 2017; Gerstenberg & Icard, 2019; Kominsky et al., 2015; Kominsky & Phillips, 2019; Henne, Pinillos & De Brigard, 2017; Henne et al., 2019; and especially Morris et al., 2019; Quillien & Barlev, 2021)[8].

Additionally, some counterfactuals come to mind more readily than others. Intuitively, if you witnessed the event leading to the forest fire, your first thought would probably not be "what if there had been no oxygen in the air?", because this possibility is extremely unlikely *a priori*. Instead you probably would be thinking about the fact that lightning might not have struck.

Across the counterfactuals that people spontaneously generate, the correlation between "there is oxygen in the air" and "the forest is on fire" is very low (notably because of the many counterfactuals where there is oxygen but nothing to spark the fire). This fact explains why it feels strange to say that oxygen caused the fire. By contrast, lightning bolts and forest fires tend to strongly co-occur across counterfactuals, and therefore we intuitively say that the lightning bolt caused the forest fire (Quillien, 2020).

Counterfactual models of causation also successfully predict that normative considerations impact causal intuitions. People are biased to generate counterfactuals that are

[8] For ease of exposition we are somewhat simplifying the theory. Obviously, a correlation between two variables is not always indicative of causation -- for instance, lightning causes both thunder and fire, so there will be a high correlation between thunder and fire across counterfactual worlds, but it would be invalid to judge that the thunder caused the forest fire. The model developed in Quillien (2020) easily deals with such cases, but getting into these details is not crucial here.

statistically normal (as explained above) but they are also biased towards counterfactuals that are *prescriptively* normal, i.e. where agents do not violate ethical or legal norms (Byrne, 2016). For instance, if two cars collide at an intersection, it feels more natural to ask "what if the car that went through the red light had stopped instead" rather than mentally changing the behavior of the car that went through the green light. Therefore, we are more inclined to say that the car that went through the red light caused the collision (see Hitchcock & Knobe, 2009; Samland, Josephs, Waldmann, & Rakoczy, 2016; Icard et al., 2017; for the thesis that *normality* has both a descriptive and prescriptive meaning, see Kahneman & Miller, 1986; Bear & Knobe, 2017; Bear et al., 2020; Phillips, Morris & Cushman, 2019).

In sum, the psychology of causal judgment is no longer an entirely black box to cognitive scientists. Recent models naturally explain why people often tend to deny a causal role to events that would be considered causal under the egalitarian conception of causation prevalent in philosophy. Additionally, counterfactual models of causal judgment can make (often fine-grained) predictions about the causal intuitions that people will have in a given situation. If our theory of intentional action is correct, then the variables that these models identify as important to causal judgment should also shape intentionality judgments. We will test this prediction in studies 2, 4 and 5.

Recall that earlier we stated our hypothesis as:

For the human mind, an agent did X intentionally if the agent's attitude toward X caused X, and caused X according to the typical causal model implicit in our commonsense psychology.

We are now able to specify that by "caused", we mean the intuitive, graded concept of causation, instead of the egalitarian notion. Next we explore the fit between our account and people's intuitions.

4) Explaining intuitions about intentional action

4.1) Deviant causation

We first address cases in which people have the kinds of intuition that seem *a priori* most damning to our causalist account. In cases of "deviant causation", an agent's attitude toward X caused X, but intuition suggests that the agent did not intentionally do X. We suggest that these cases come in two kinds.

The first kind of case is where an agent's attitude causes an outcome in a way that deviates from the domain-specific causal model of commonsense psychology. In such situations, our account explicitly predicts that people will not attribute intentionality. The **King** scenario in section 3.1 is one such example. Another case was famously discussed by Donald Davidson:

> **Climber.** "A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never chose to loosen his hold, nor did he do it intentionally" (Davidson, 1980).

We suggest that people do not judge the climber's action as intentional because the event cannot be represented using the standard causal model depicted in Diagram 1b. Under this causal

model, the only way the climber's desire can cause him to loosen his hold is by affecting the expected utilities he computes for each alternative course of action, altering the decision he eventually makes. But this is not what happens in **Climber**. In order to represent the event, we need to use an 'augmented' causal model, namely the one depicted in Diagram 3, which includes a new causal path involving nervousness. This alternative causal pathway prevents people from judging that the climber's desire (i.e. his positive attitude toward the outcome) caused the outcome in the right way, and therefore it makes them reluctant to judge the event as intentional[9].

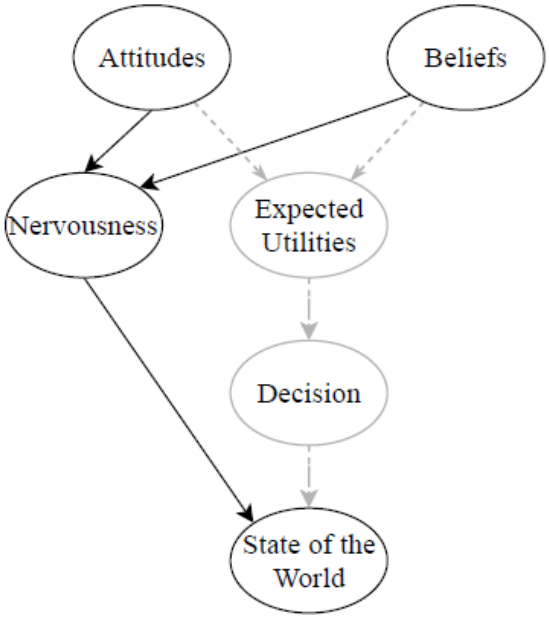

Diagram 3: Causal model for the **climber** scenario. Attitudes affect the state of the world via a causal pathway (black solid arrows), which is different than the one pre-specified by commonsense psychology (grey dashed arrows).

[9] Cosmides (1985, chapter 5) makes a similar argument to account for cases of causal deviance in social exchange.

The second kind of case stems from the fact that people's domain-general concept of causation is not egalitarian (recall that, e.g., it feels strange to say that oxygen in the air caused the forest fire). As a result, people are sometimes reluctant to judge that an agent's attitude caused an outcome, even when the outcome counterfactually depended on the attitude.

For example, Knobe (2003b) asked participants about the following case:

> **Bull's-eye.** Jake desperately wants to win the rifle contest. He knows that he will only win the contest if he hits the bull's-eye. He raises the rifle, gets the bull's-eye in the sights, and presses the trigger. But Jake isn't very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild . . . Nonetheless, the bullet lands directly on the bull's-eye. Jake wins the contest.

Only 28% of participants judged that Jake intentionally hit the bull's eye (by contrast, 79% of participants ascribed intentionality to Jake when he was described as an expert marksman executing a perfect shot)[10]. Under our account "Jake intentionally hit the bull's eye" means, roughly: "The bullet hit the bull's eye because Jake wanted it to". The latter statement is true according to an egalitarian theory of causation: the bullet would not have hit the bull's eye if Jake had not wanted it to. However, we suspect that people would not share the verdict of the egalitarian theory. Intuitively, the real cause of Jake's success was dumb luck.

Why does it feel strange to say that the bullet hit the bull's eye because Jake wanted it to? The counterfactual model of causal judgment (Quillien, 2020) described in section 3.2 provides an explanation. When people think about the case, they implicitly generate several

[10] Interestingly, in a scenario pair which is similar, except that Jake's intention is immoral, causal deviance has much weaker effects on judgments of intentionality (Knobe 2003b; see also Sosa, Holbrook & Swiney, 2015). We return to this point in the General Discussion.

counterfactuals to the event. Across these counterfactuals, they compute the correlation between "Jake wants to hit the bull's eye" and "the bullet hits the bull's eye". This correlation is low: Jake is a novice marksman, so in most counterfactuals, he wants to hit the target but completely misses. Therefore, Jake's attitude is not an important cause of the outcome, and Jake did not intentionally hit the bull's eye.

4.2) Explaining recurrent features in existing accounts

There are many different accounts of the meaning of "intentionally", but historically most of them have shared the two following requirements. In order for an action to be intentional, the agent must have a Desire for the outcome to occur and a Belief that their action will bring about this outcome. When asked for their explicit definition of the concept, laypeople also systematically say that Desire and Belief are central to intentional action (see Malle & Knobe, 1997, for a historical review, and for empirical data about the explicit folk concept).

By contrast, our account does not explicitly mention either Desire or Belief as being necessary for an action to be intentional. Nevertheless, we show below that it can explain why in most cases people will tend to view these features as essential to intentionality. Our account also predicts that people will sometimes attribute intentionality to agents who do not have a desire for the outcome, or to agents who have only a weak belief that their action would bring about the outcome. As we will show, such situations do occur.

4.2.1) ***Desire is important, but not essential.***

A desire for X is simply a positive attitude toward X. Attitudes toward an outcome are more likely to lead to that outcome when they are positive. For instance, if someone eats vanilla

ice cream, it is usually because they wanted to eat vanilla ice cream. Therefore, the typical case when an attitude leads to X is when the agent has a desire for X. Ergo, under our account, in most cases agents do X intentionally because they want X to happen.

Yet even a neutral, or a negative attitude toward X can cause X. Notably, a negative or neutral attitude can be considered causal if it is *not negative enough*. Consider the well-known Chairman vignette designed by Knobe (2003a):

Chairman. The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.' The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was harmed.

When reading the story, we infer that the chairman's attitude toward harming the environment is neutral, or perhaps mildly negative: the chairman might view harming the environment as a somewhat unfortunate outcome, but not enough so as to overcome his lust for profit. But the normative expectation is that the chairman should be sufficiently opposed to harming the environment that he would refuse to implement the program. Therefore we tend to mentally replace the event with one where the chairman's attitude conforms to this normative expectation (we are unlikely to mentally replace the event with one where the chairman values the environment less than he does in the actual situation). In such a counterfactual, changing the chairman's attitude prevented harm to the environment.

Thus, computational accounts of causal judgment (Icard, Kominsky & Knobe, 2017; Quillien, 2020) predict that people will judge that the environment was harmed *because* the

chairman did not care about the environment (we test this prediction in study 4). In turn, this predicts that people will judge that the chairman intentionally harmed the environment. This is indeed what Knobe (2003a) finds: people judge that the chairman intentionally harmed the environment, contradicting the standard assumption that only agents who want X to happen can do X intentionally.

4.2.2) **Belief is important, but can be weak.**

Our account implies that for X to be intentional, the agent must believe that his action has a non-zero likelihood of bringing about X. This requirement follows from the causal model implicit in commonsense psychology. In this causal model, attitudes can only influence the state of the world by affecting expected utilities, and attitudes can only influence expected utilities if the agent has the relevant beliefs. Let us say I have a choice between doing A and not doing A. If I believe that regardless of what I choose, X will not happen, then my attitude toward X cannot possibly have the right kind of causal influence on whether I choose to do A[11]. Therefore I must believe that doing A will lead to X with non-zero probability in order to do X intentionally.

Note that this leaves open the following possibility: an agent may be intentional even if he believes that his action has a very low probability to lead to X. What matters is that the agent believes that his action increases the probability of X *at least a little*. For instance, if the agent has a sufficiently strong desire for X, and/or taking the action is not very costly, then they may decide to take the action *because* they think it may lead to X, even if they think that this probability is very dim.

[11] More generally, if I believe that the probability that X will happen is the same regardless of whether I choose A or not A, then my attitude toward X cannot play a causal role in my decision.

Intuition seems consistent with this possibility. For instance, Davidson (1980) remarks: "in writing heavily on this page I may be intending to produce ten legible carbon copies. I do not know, or believe with any confidence, that I am succeeding. But if I am producing ten legible carbon copies, I am certainly doing it intentionally" (Davidson, 1980, Essay 4, p.82). We test this intuition more systematically in study 3.

In summary, our account can explain why people do not attribute intentionality to agents in cases of deviant causation, and why standard theories of intentional action have the features that they do. Our account can also explain why people's intuitions sometimes deviate from the predictions made by their explicit theory of the concept and by standard scholarly theories.

We now turn to empirical tests of our definition.

5) Empirical tests

Here we report the results of six empirical tests of our account. Our general strategy is relatively simple: if people's judgments of intentionality derive from their mental representations of causation, then their judgments of causation and their judgments of intentionality should track each other. In other words, when people judge that an agent did X intentionally, they should also judge that X was causally dependent on the agent's attitude toward X. As a corollary, manipulations that are known to affect causation judgments should affect intentionality judgments, and vice-versa.

Note that our account does not *strongly* predict that causation and intentionality judgments will always perfectly track each other. We assume that when people compute whether X was caused "in the right way" for the purpose of assessing intentionality, they rely on two

kinds of cognitive systems: the domain-specific model implicit in commonsense psychology, and the domain-general concept of causation. By contrast, when people are explicitly asked whether the agent's attitude toward X *caused* X, it is possible that they rely more exclusively on their domain-general concept of causation. If this is the case, then one might construct contrived thought experiments, similar to **King** (see section 3.1) where people judge that the agent's attitude toward X was highly causal, even though they judge that he did not do X intentionally. We test this prediction in study 6. More generally, a variety of pragmatic and motivational factors may distort how people answer queries about intentionality and causation, leading to non-identical patterns of responses.

Nonetheless, the prediction that causation and intentionality judgments will tend to track each other constitutes a non-trivial prediction of our account, which makes it worth testing. Some of our studies (studies 2-4) were additionally designed to provide empirical support to the explanations we have given for some phenomena in the previous sections of this paper. In summary, the studies we report are meant to provide evidence for a causalist definition of "intentionally", and to serve as a proof of concept that cognitive science models of causal cognition can shed light on people's intuitions about intentional action. Data and R code for all studies are available at the Open Science Framework at https://osf.io/42x7h/?view_only=64e21726c536419c9942ac2ca1aca9c1.

5.1) Study 1: intuitions in everyday situations

Study 1 was a very simple preliminary test of our theory. We asked participants to read a series of 19 short statements about various events involving a person called Anne. Half the

participants were asked to rate whether Anne was doing what she was doing intentionally. The other half were asked whether what happened depended on whether Anne wanted it to happen. We predicted that answers to both questions would closely track each other.

5.1.1) Methods

Participants. We recruited 200 participants from AmazonMTurk. Five participants were excluded from analysis for failing a catch item (typing "4" in response to a picture displaying the question "What is 12-8?" – we used this catch item in all studies reported here), leaving a total of 195 participants (97 female).

Stimuli. Participants read a series of 19 short sentences, which we adapted from study 1 in Malle & Knobe (1997). All sentences describe a person called Anne doing something, for instance "Anne was sweating", "Anne got admitted to Princeton", "Anne stole a pound of peaches", etc (see Appendix for complete list of statements). The sentences were identical to the statements used in the original study, except that we modified some of them so that they were all in the past tense. These stimuli were originally designed by Malle & Knobe (1997) for a different purpose than the present study, namely to study inter-rater agreement in intentionality ratings. Conveniently, they were designed so that ratings would span a wide range, with different stimuli expected to elicit low, intermediate and high intentionality ratings.

Procedure. All 19 statements were presented on the same page, in random order. Participants were randomly assigned, in a between-subjects design, to either an "Intentionality" or a "Dependence" condition. Participants in the "Intentionality" condition were asked to rate, for each statement, whether Anne did what she did intentionally, on a likert scale from 1 (not intentional at all) to 7 (very intentional). Participants in the "Dependence" condition were asked

to rate, for each statement, whether the event described depended on whether Anne wanted it to happen or not, on a scale from 1 (completely independent) to 7 (completely dependent). We asked about dependence instead of explicitly using the expression ‘causally depended’ because the latter sounds less natural, and in the context of our statements it is clear that ‘dependence’ refers to a causal link[12].

[12] More generally, across the studies reported here we could not systematically use the same measure of how much the participants judged that the agent’s attitude toward X caused X, because the abstract concept of “attitude toward X” is difficult to express in English. So, for instance, in studies where the situation makes it clear that the agent had a desire for X, we ask about whether X happened because the agent wanted X.

5.1.2.) Results and discussion.

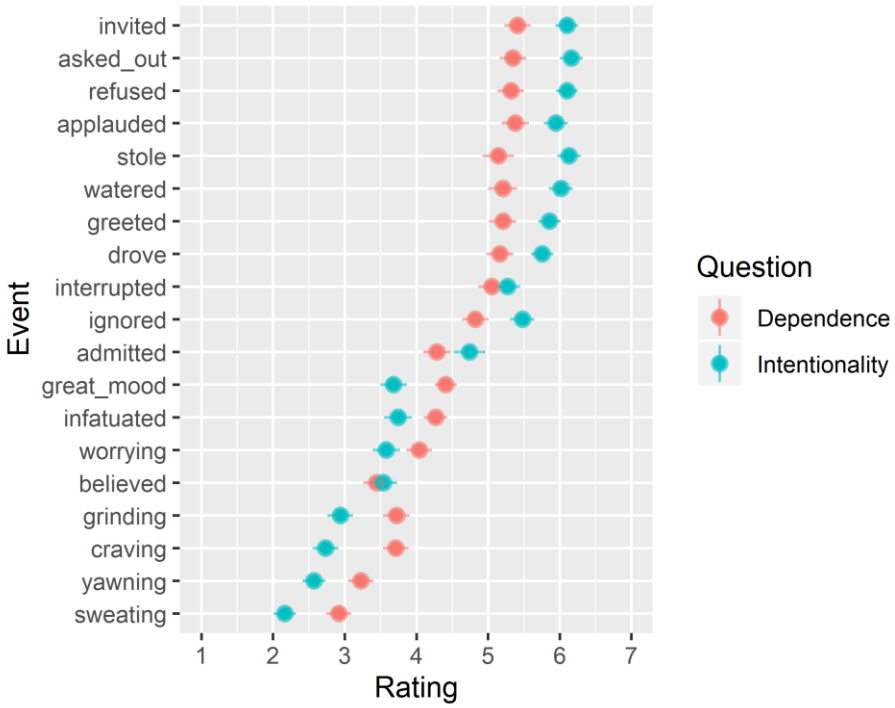

Figure 1: mean ratings for Dependence and Intentionality, for each sentence in Study 1. Error bars represent standard errors of the mean.

The item-level correlation between causal dependence and intentionality ratings was almost perfect, $r(17) = .96$, $p < .001$; see Figure 1.

Interestingly, intentionality ratings look like "stretched out" versions of the dependence ratings: they are more likely to lie close to the endpoints of the scale. We do not really know why this is the case. It may be that, compared to causation, people are more reluctant to treat intentionality as a graded concept. Or maybe participants were slightly more confused by the causal dependence question.

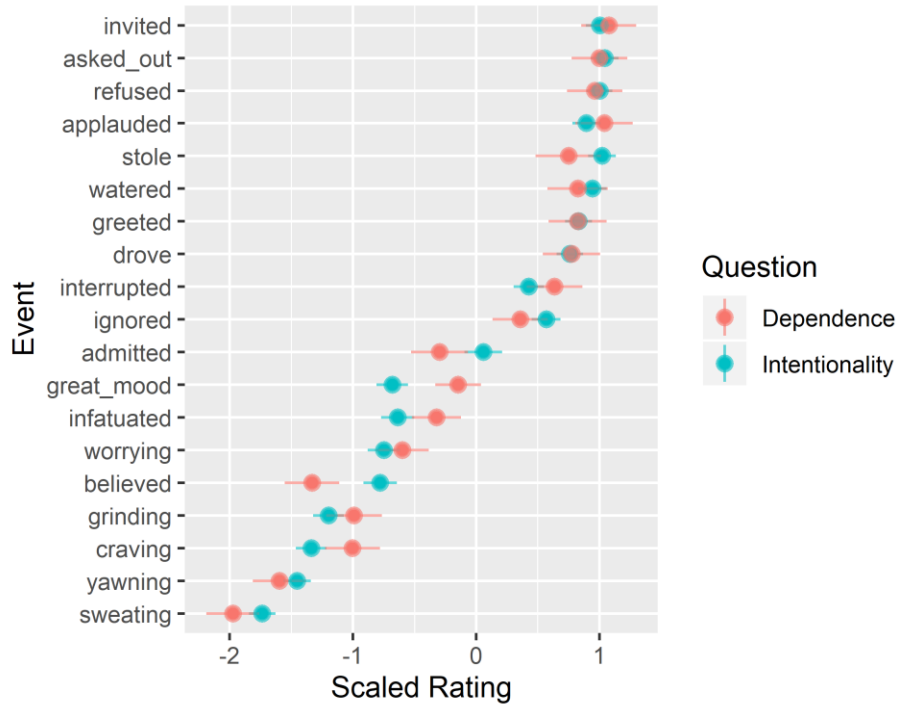

Figure 2: mean scaled ratings for dependence and intentionality, study 1. Error bars represent standard errors of the mean.

To get a better subjective sense of the tight fit between ratings on the two questions across events, we can look at scaled mean ratings: we created Figure 2 by computing z-scored means for dependence and intentionality ratings independently (by, e.g., subtracting the mean intentionality rating for a given event from the grand mean rating for intentionality, and dividing by the standard deviation in mean intentionality ratings across events).

The results strongly support the idea that people understand intentionality in terms of the causal dependence of the outcome on the agent's attitude toward the outcome. However, the sentences we used depict everyday situations. In the following studies, we "stress-test" our theory by exposing it to more exotic cases.

5.2) Study 2: Causal Deviance

On the surface, cases of 'causal deviance', where an agent's attitude toward X caused X, yet the agent did not do X intentionally, seem like obvious counter-examples to our theory.

We have argued that in many such cases, causation occurs only in the egalitarian sense of metaphysicians; to our mind's intuitive concept of causation, the attitude did not really cause X, or did so only weakly – hence the intuition that the agent did not do X intentionally. For instance, a typical case of causal deviance (inspired by Chisholm, 1966) goes like this: Alice wants to kill Bob and decides to drive to the gun store; while driving, she runs over a pedestrian, who turns out to be Bob. This case elicits the intuition that Alice did not intentionally kill Bob. Yet, according to the egalitarian notion of causality, Alice's desire to kill Bob caused Bob to die.

This case can be reconciled with our account by positing that people view "Alice wants to kill Bob" as only weakly causal for "Bob died". Counterfactual models of causal judgment (Quillien, 2020; Icard et al., 2017; see also Kominsky et al., 2015) predict that people will indeed assign a relatively low causal weight to Alice's desire (i.e. to her positive attitude toward the outcome). The fact that Bob happened to cross the street right at this particular moment is a coincidence, i.e., an event with low *a priori* probability. It is easy to think of counterfactuals where Alice wants to kill Bob and drives toward the gun store, but Bob does not cross the street, or does so at a slightly different time. In these counterfactuals, Bob is still alive right after Alice drives through that particular street. This means that, across possible counterfactuals to the event, there is a relatively low correlation between "Alice wants to kill Bob", and "Bob dies". As a consequence, people will be reluctant to judge that Alice's desire to kill Bob was the cause of Bob's death. They should then also deny that Alice intentionally killed Bob.

In study 2, we test this prediction by probing people's intuitions about causation and intentionality in the "causal deviance" case described above. We also created a matched story where the causal link between Alice's desire and Bob's death is straightforward. We predict that compared to the straightforward causal link story, the causal deviance story will elicit lower intentionality *and* lower causation ratings.

5.2.1) Methods

Participants. We recruited 203 participants from Amazon MechanicalTurk. Two participants were excluded from analysis for failing a catch item, leaving a final sample of 201 participants (112 female, 1 other).

Stimuli and Procedure. Participants were randomly assigned (between-subjects) to one of the two following vignettes:

> Alice hates Bob. One day she decides to go buy a gun, in order to kill him. She gets in her car and starts driving in the direction of the gun shop. Someone suddenly crosses the street in front of her. [She realizes that the pedestrian is Bob. Seizing the opportunity, she steps on the gas and runs him over / Unbeknownst to her, the pedestrian is Bob. She steps on the brake, but it is too late and she runs him over]. Bob dies on the spot.

Each participant was asked two questions: an Intentionality and a Causation question. Question order was randomized across participants: half the participants answered the Intentionality question first, the other half answered the Causation question first. Questions appeared on different pages of the computer-based survey. The Intentionality question was the same for every participant: they were asked to rate their agreement with the statement "Alice intentionally killed Bob" on a 1-7 likert scale (1:strongly disagree, 7:strongly agree). For

exploratory purposes, we varied the wording of the Causation question across participants: half the participants were asked to rate their agreement with the statement "Alice's desire to kill Bob caused Bob to die", the other half were asked to rate their agreement with the statement: "Bob died because of Alice's desire to kill Bob".

5.2.2) Results

Results were consistent with our predictions (see Figure 3). Intentionality ratings were higher in the Normal condition (M=6.78, SD=.80) than in the Deviant condition (M=2.76, SD=1.89), $t(128.95) = 19.03$, $p < .001$, $d = 1.89$. Similarly, Causation ratings were higher in the Normal condition (M=6.17, SD=1.52) than in the Deviant condition (M=3.48, SD=2.10), $t(175.9) = -10.38$, $p < .001$, $d = 1.49$.

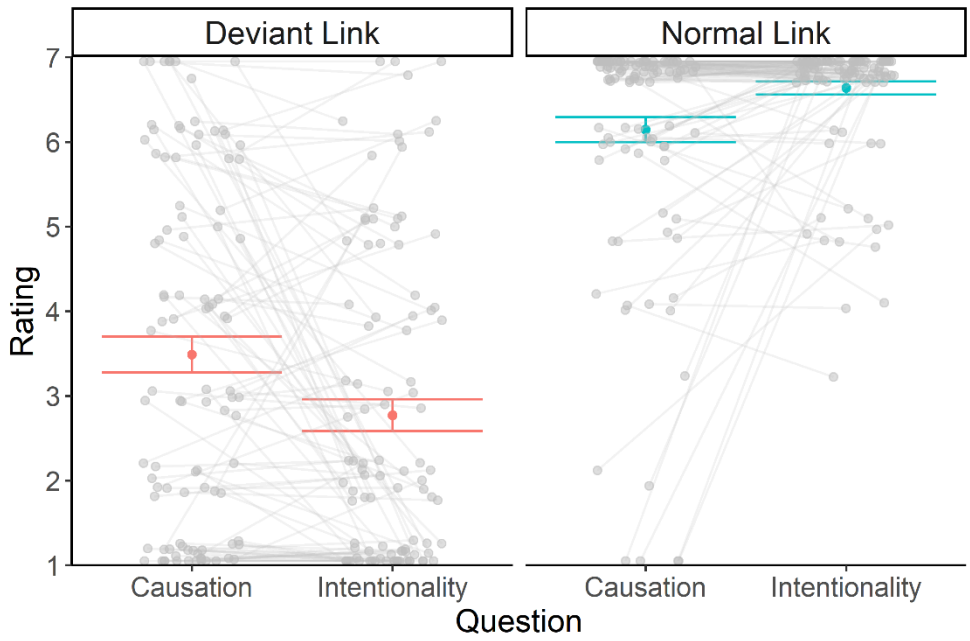

Figure 3: Ratings of Causation and Intentionality for Normal and Deviant causal links. Error bars represent standard errors of the mean. Individual data points are jittered for better visibility.

In order to compare the size of these effects, we conducted a mixed Anova, which revealed that the type of causal link (Normal vs Deviant) had a larger effect on intentionality than on causation ratings (interaction: $F(1, 199) = 27.10$, $p < .001$, $\eta^2_{partial} = .12$).

We did not find any order effects or wording effects. The effect of the type of causal link on Causation ratings did not depend on the wording of the causation question ("caused" vs "because")[13] , $F(1,197) = 0.00$, $p = .95$; and the order of questions did not affect the effect of causal link on either the causation or the intentionality rating (all $Fs < .85$, all $ps > .35$).

We also tested whether the effect of the type of causal link on intentionality ratings was mediated by causation ratings. To do so, we conducted a mediation analysis with 5000 resamples using the lavaan package in R (Rosseel, 2012). We found that there was a significant indirect effect of condition on intentionality ratings via causation ratings (95% CI [0.642, 1.561]; see Supplementary Information at https://osf.io/dp5xr/?view_only=64e21726c536419c9942ac2ca1aca9c1 for the full model).

5.2.3) Discussion

In both versions of the story we used, the egalitarian answer to the causation question is that Alice's desire to kill Bob caused Bob to die: if Alice had not wanted to kill Bob, she would not have taken the wheel, and Bob would still be alive. Yet people's causal intuitions were more subtle: people viewed Alice's desire as causally important to Bob's death in the scenario where the causal link was straightforward, but viewed it as much less so in a "causally deviant" scenario where counterfactual models of causal judgment assign low causal strength to Alice's

[13] We are not aware of much research that has looked at the differences between 'cause' and 'because' statements, although see Livengood & Machery (2007) for a preliminary investigation in the context of causation by absence.

desire. Correspondingly, they only considered that Alice killed Bob intentionally in the scenario involving a non-deviant causal link.

Though our manipulation of the causal link had a similar effect for both causation and intentionality ratings, this effect was somewhat stronger for intentionality than for causation. We do not really know why this is the case, but we note that this is consistent with the pattern found in Study 1: compared to causation ratings, average intentionality ratings are closer to the endpoints of the scales.

Overall, the results of study 2 suggest one way that cases of causal deviance can be consistent with a causalist account of "intentionally". Even when the agent's attitude toward the outcome was technically necessary for the outcome, people may not be judging it as strongly causal. If, across counterfactuals to an event, the outcome was only weakly causally dependent on the agent's attitude, people will tend to deny that the agent's attitude was the cause of the outcome. As a result, they will deny that the event was intentional.

In sum, the logic of our intuitive, domain-general concept of causation might explain many cases of causal deviance. Our account also predicts that in other cases, a causal link may be treated as deviant because it travels outside of the causal model of commonsense psychology. We examine this possibility in a later study (study 6).

5.3) Study 3: Intentionality despite weak belief

A standard feature of virtually all accounts of intentional action is that an agent who does X intentionally must believe that their action will lead to X (see section 2). As we explain in section 4.2, our account shares this feature, but makes the additional prediction that there will be

cases where an agent is judged as intentional even if their belief that the action will lead to the outcome is very weak.

A study by Mele & Cushman (2007) provides some support for the prediction. Participants in that study read the following vignette:

Bowling. Earl is an excellent and powerful bowler. His friends tell him that the bowling pins on lane 12 are special 200-pound metal pins disguised to look like normal pins for the purposes of a certain practical joke. They also tell him that it is very unlikely that a bowled ball can knock over such pins. Apparently as an afterthought, they challenge Earl to knock over the pins on lane 12 with a bowled ball and offer him ten dollars for doing so. Earl believes that his chance of knocking over the pins on lane 12 is very slim, but he wants to knock them down very much. He rolls an old bowling ball as hard as he can at the pins, hoping that he will knock down at least one. To his great surprise, he knocks them all down! The joke, it turns out, was on Earl: The pins on lane 12 were normal wooden ones. (Mele & Cushman, 2007, p. 187)

Participants overwhelmingly agreed that Earl intentionally knocked down the pins (M=6.36 on a 1-7 scale), even though the story mentions that he believes his chance of doing so is very slim.

We designed Study 3 to test the generalizability of this finding, and to confirm that in this sort of case, people judge that the agent only has a very weak belief that his action will lead to the outcome. We asked participants to read Mele & Cushman's **Bowling** scenario; in addition to their ratings of intentionality, we also asked them whether the agent believed that his action would lead to the outcome, and whether the outcome was caused by the agent wanting the

outcome to occur. We also designed two additional vignettes that we predicted would elicit high intentionality ratings despite low belief ratings.

5.3.1) Methods

Participants. We recruited 90 participants on Amazon MechanicalTurk. Twenty-four participants were excluded from analysis for failing a catch and/or a comprehension item (see below), yielding a final sample of 66 participants (32 female, 1 who declined to state).

Stimuli and Procedure. Participants were randomly assigned (between-subjects) to one of three vignettes. The **Bowling** vignette was adapted verbatim from Mele & Cushman (2007). The **Sabotage** vignette depicted a disgruntled worker in a power plant who tries to shut down the main reactor by pushing a red button, even though he knows he does not have the necessary security key – because of an oversight, the safety feature was turned off, and by pushing the red button he shuts down the reactor. In the **Shooter** vignette, a shooter realizes at the last moment that he forgot to put bullets in his rifle, but decides to fire at his victim anyway – as it turns out, there were actually some bullets left over in the rifle, and the victim dies. See appendix for the full text of the vignettes.

Each participant was asked three questions: an Intentionality, a Belief, and a Causation question. Since our main hypothesis was about intentionality and belief, the causation question was always presented last. Half the participants answered the Intentionality question first, the other half answered the Belief question first. Questions appeared on different pages of the computer-based survey. Participants were asked to rate their agreement with the following statements, on a 1-7 likert scale (1: strongly disagree, 7: strongly agree):

-Earl intentionally knocked down all the pins

-Earl believed that he would knock down all the pins

-All the pins were knocked down because Earl wanted to knock down all the pins

The first page also featured a Comprehension question (e.g. “Earl’s friends didn’t know that the pins on lane 12 were normal wooden ones”, with options: True/False/Impossible to tell); participants failing to provide the correct answer were excluded from analysis.

(See Appendix for the questions associated with the other two vignettes).

5.3.2) Results and discussion

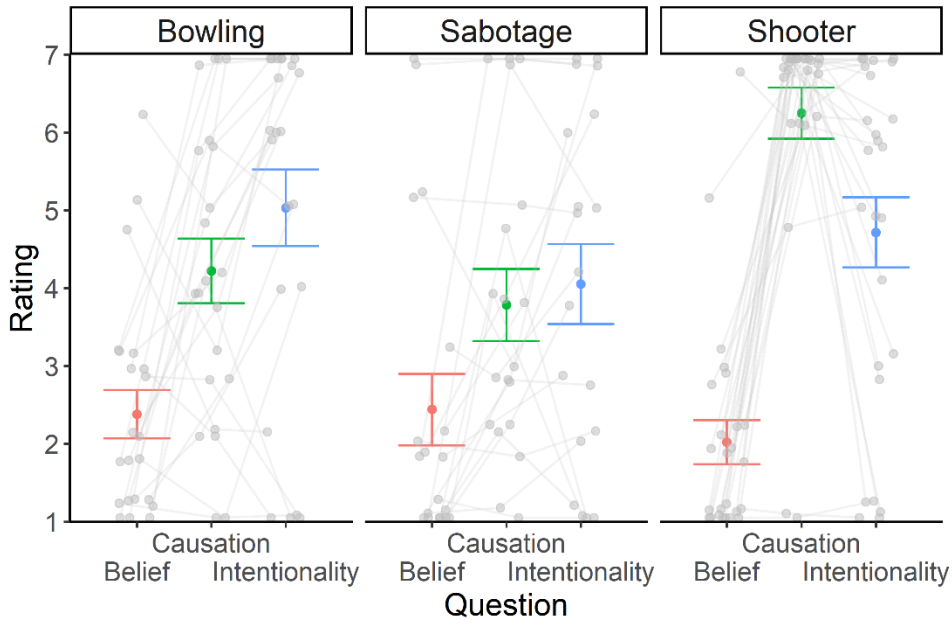

Figure 4: Belief, causation and intentionality ratings for each vignette. Error bars represent standard errors of the mean. Individual data points are jittered for better visibility.

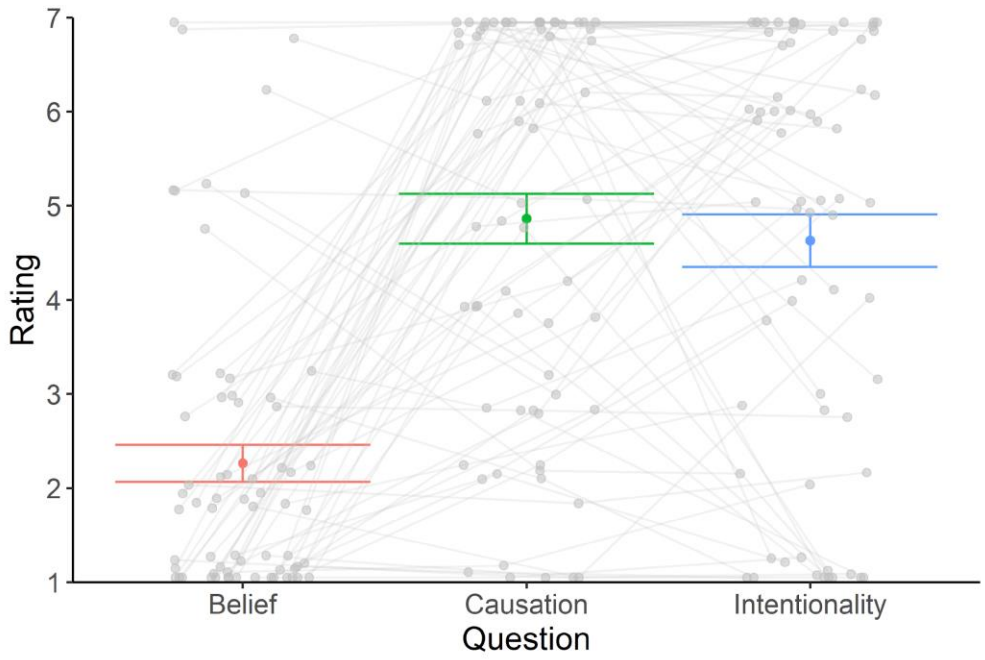

Figure 5: Belief, causation and intentionality ratings, collapsed across vignettes. Error bars represent standard errors of the mean. Individual data points are jittered for better visibility.

Participants tended to attribute weak belief to the agent (M = 2.24, SD = 1.63), although they attributed medium-to-high amounts of intentionality (M = 4.64, SD = 2.30) and causation (M = 4.88, SD = 2.17), see Figures 4-5.

Intentionality ratings were higher than Belief ratings, as assessed by a 2*3 mixed Anova with Question Type (Intentionality vs Belief) and Vignette as predictors: the main effect of Question Type on ratings was significant, $F(1,63) = 69.6$, $p < .001$, $\eta^2_{partial} = .51$. There was no interaction, $F(2,63) = 1.43$, $p = .25$, and no main effect of Vignette, $F(2,63) = .47$, $p = .63$.

We found no effect of Question Order (whether participants saw the Intentionality or the Belief question first) on ratings: a 2*2 Anova with Question Type and Question Order as

predictors failed to find a main effect of Question Order, $F(1,128) = 1.00$, $p = .32$, or an interaction between Question Type and Question Order, $F(1,128) = 1.27$, $p = .26$.

Results confirm that a strong belief that the action will lead to the outcome is not a necessary requirement for intentionality. According to our account, this is because some attitudes can be strong enough that they are considered to have causal power even in the absence of a strong belief that the action will lead to the outcome. Consider for instance the **Shooter** vignette. For most people, even a remote possibility that there might be bullets left in the rifle is enough to prevent them from shooting an unloaded rifle at someone. Someone who does so anyway demonstrates an abnormally low valuation of the target's life. Because it is so abnormal, this low valuation is readily selected as the cause of the victim's death. Therefore the shooter intentionally killed the victim.

We also observe that overall, causation ratings were close to intentionality ratings. Study 3 was not explicitly designed to test this prediction (as causal judgments were always elicited last), so this result should be interpreted with caution, but it is one additional piece of evidence that causation is what matters for intentionality, even in the realm of somewhat contrived thought experiments.

In the next study, we turn to what is arguably the most famous of these thought experiments.

5.4) Study 4: a side-effect effect for causality.

In section 4.2, we introduced the Knobe's (2003a) **chairman** vignette as an example of a case where an agent is judged to do X intentionally even though he does not desire X. Knobe

also found a striking asymmetry in people's judgments: participants reading the same vignette, with the word "harm" replaced by "help" overwhelmingly denied that the chairman intentionally helped the environment. There have been a large number of attempts to explain this asymmetry, known as the Knobe effect or side-effect effect (Nadelhoffer, 2006; Wright & Bengson, 2009, Pettit & Knobe, 2009; Uttich & Lombrozo, 2010; Cova, Dupoux & Jacob, 2012; Adams & Steadman, 2004; Hindricks, 2014; Sripada, 2012; Sloman, Fernbach & Ewing, 2012; Machery, 2008; Leslie, Knobe & Cohen, 2006).

The side-effect effect makes intuitive sense in our framework: in the Harm vignette, it seems natural to say that the environment was harmed because the chairman does not care about the environment, while in the Help vignette, it seems unnatural to say that the environment was helped because the chairman does not care about the environment.

This asymmetry in causal attributions can be explained by the psychological logic of causal judgment. We assume that there is a normative expectation that people ought to strongly value the environment. Therefore, the attitude of the chairman toward the environment is abnormal. When they make causal judgments, people will tend to generate counterfactuals where the chairman has a more normal attitude, i.e. counterfactuals where the chairman values the environment more than he does in the vignette.

In the Harm case, considering these counterfactuals ends up changing the outcome: in counterfactuals where the chairman values the environment sufficiently highly, he will oppose the program. By contrast, in the Help case, whether the chairman has a neutral or a positive valuation of the environment does not matter for the outcome (in both cases the chairman implements the program, and the environment is helped). Therefore, when people consider different attitudes that the chairman could have had, they see that this has a large effect on the

outcome in the Harm case, but a very small effect in the Help case. As a result, they judge that the chairman's attitude is much more causally important in the Harm case.

Study 4 was designed to test the prediction that people make higher causal attributions in the Harm case than in the Help case.

5.4.1) Methods

Participants. We recruited 210 participants from Amazon MechanicalTurk. Twelve participants were excluded from analysis for failing a catch item, yielding a final sample of 198 participants (83 female).

Stimuli and Procedure. Participants were randomly assigned (between-subjects) to read one of the following vignettes, adapted from Knobe (2003a):

"The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also [harm/help] the environment.' The chairman of the board answered, 'I don't care at all about the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was [harmed/helped]."

In both conditions, participants were asked a Causation question first, and then an Intentionality question on a separate page. They were asked to rate their agreement with the following statements, on a 1-9 likert scale (1: strongly disagree, 9: strongly agree):

- The fact that the chairman does not care about the environment caused the environment to be [harmed/helped].

- The chairman intentionally [harmed/helped] the environment.

5.4.2) Results

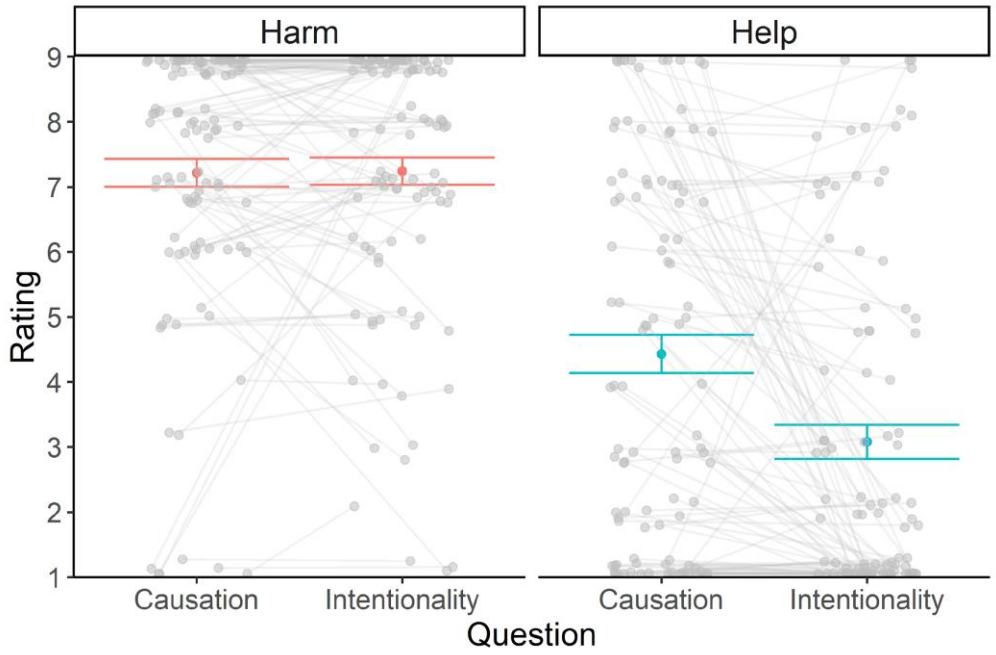

Figure 6: Causation and intentionality ratings as a function of condition, Study 4. Error bars represent standard errors of the mean. Individual data points are jittered for better visibility.

As predicted, we found a side-effect effect for causation judgments: causation ratings were higher in the Harm (M=7.23, SD=2.16) than in the Help case (M=4.42, SD=2.92), $t(180.55) = 7.69$, $p < .001$, $d = 1.10$. We also replicated the original side-effect: intentionality ratings were higher in the Harm (M=7.26, SD=2.09) than in the Help case (M=3.06, SD=2.61), $t(187.04) = 12.5$, $p < .001$, $d = 1.79$.

To compare the size of these effects, we conducted a 2*2 mixed Anova with Question Type (intentionality vs causation) and Condition (Harm vs Help) as predictors. We found a

significant interaction, $F(1, 196) = 15.87$, $p < .001$, $\eta^2_{partial} = .07$, showing that the effect of Condition on ratings is higher for intentionality than causation.

We also tested whether the effect of condition on intentionality ratings was mediated by causation ratings. To do so, we conducted a mediation analysis with 5000 resamples using the lavaan package in R (Rosseel, 2012). We found a significant indirect effect of condition on intentionality ratings via causation ratings (95% CI [0.80, 1.93]; see Supplementary Information at https://osf.io/dp5xr/?view_only=64e21726c536419c9942ac2ca1aca9c1 for the full model).

5.4.3) Discussion

Results of Study 4 are consistent with a causalist account of the side-effect effect. On the other hand, these results are not completely surprising: analogues of the side-effect effect have been found for many other types of judgments, such as judgments of whether an agent knows something or is in favor of something (Pettit & Knobe, 2009; Beebe & Buckwalter, 2010). As such, one could also have predicted the existence of a causal side-effect effect on a completely atheoretical basis, simply by generalizing from these already known similar effects.

Therefore, a challenge for our account is to show that the explanation we give for the side-effect effect can also be used to make genuinely novel predictions: predictions that would not follow from other accounts of the effect, or from simple generalization. We take on this challenge in the next study.

5.5) Study 5: Reversing the effect of norm violation on intentionality

Many accounts of the side-effect effect rely on the idea that people attribute higher intentionality to agents who violate a norm (Nadelhoffer, 2006; Pettit & Knobe, 2009; Hindricks, 2014; Uttich & Lombrozo, 2010; Holton, 2010; Alicke, 2008). For instance, people may make such attributions because of a motivation to blame the norm violator (Nadelhoffer, 2006; Alicke, 2008), or because one can make stronger mental state inferences about someone who violates a norm (Uttich & Lombrozo, 2010).

Our account also relies on the notion of norm violation, but makes more subtle predictions. It holds that norm violations have an effect on intentionality judgments because they have an effect on causation judgments. Causation judgments are sensitive to norm violations because they rely on counterfactuals, and people sample counterfactuals as a function of how normal they are – where 'normal' has a broad meaning, encompassing statistical, normative and functional considerations.

As such, our account predicts that in many cases, norm violators (e.g. agents who engage in immoral actions, violate a conventional norm, behave in a different way than they usually do, etc) will be judged as more intentional than non-norm-violators. However, it also predicts the existence of cases where this effect *reverses*: cases where norm violators are judged as *less* intentional than non-norm-violators.

Specifically, computational models of causal judgment (Icard, Kominsky & Knobe, 2017; Quillien, 2020) predict the following interaction between normality and causal structure (see SI at https://osf.io/dp5xr/?view_only=64e21726c536419c9942ac2ca1aca9c1 for an informal explanation):

-In situations that have a conjunctive causal structure (i.e. situations where several factors are jointly necessary to bring about an outcome), abnormal events are judged more causal than normal events. This predicted effect is known as *abnormal inflation*.

-In situations that have a disjunctive causal structure (i.e. situations where several factors led to an outcome, but any one of them would have been sufficient), abnormal events are judged less causal than normal events. This predicted effect is known as *abnormal deflation*.

Correspondingly, we should observe the same interaction for intentionality judgments. Study 5 was designed to test this prediction.

We asked participants to read a story (adapted from Icard et al., 2017) where a committee must vote to approve or reject a request. We manipulated the causal structure of the situation, such that in one condition, all committee members must vote Yes for the request to be approved (conjunctive causal structure), and in the other condition, the request is approved if at least one committee member votes Yes (disjunctive causal structure). We also manipulated whether the committee members violated a statistical norm, by giving background information about what the committee members *usually* do. One member, Mr A, was described as almost always voting Yes, while the other member, Mr B, was described as almost always voting No. Then we described a vote where both members vote Yes and the request is approved: in this case, Mr A is behaving normally with respect to his usual behavior, while Mr B is violating a statistical norm: even though he usually votes No, this time he is voting Yes.

We predicted that in the conjunctive causal structure, Mr B (the norm-violator) would be judged as more intentional than Mr A, while the effect would reverse in the disjunctive causal structure. We also predicted that we would find the same pattern of effects for causation

judgments, conceptually replicating previous empirical findings (Icard et al. 2017; Gerstenberg & Icard, 2019; Morris et al., 2019; Phillips & Kominsky, 2019; Henne et al., 2019).

Finally, we included a question designed to test an alternative account for the effect of norm violation on intentionality judgment. According to Uttich & Lombrozo (2010), people attribute higher intentionality to norm violators because norm violations allow stronger inferences about an agent's attitudes. Therefore, we asked participants which inferences they thought could be made about the agents' attitudes from their decisions. We hoped to find a dissociation between causation and inference judgments, which would permit a critical test between the two accounts.

5.5.1) Methods

Participants. We recruited 199 participants from Amazon MechanicalTurk. Sixty-one participants who failed either a catch item or a comprehension item (see below) were excluded from analysis, yielding a final sample of 133 participants (71 female, 1 other).

Stimuli and Procedure. We used a 2 (Causal structure) * 2 (Normality) mixed design, with Causal Structure manipulated Between-subjects, and Normality manipulated within-subjects.

Participants were randomly assigned to read one of the following vignettes:

> At a local university, a committee is in charge of evaluating new requests for funding from professors. The committee has two members, Mr A and Mr B. In order for a request to be approved, it must be that [both committee members vote / at least one committee

member votes] in favor of the request. While Mr A almost always votes Yes, Mr B is notorious for almost always voting No.

Today, the committee is examining Professor Smith's request for new computers. Although neither committee member knows Professor Smith, they both read her application carefully. Then, both committee members cast their vote at the same time. As usual, Mr A voted in favor of the request; surprisingly, Mr B also voted in favor of the request.

Since [both committee members / at least one committee member] voted in favor of the request, Prof Smith gets funding for her new computers. (adapted from Icard et al., 2017)

Participants were either asked two Intentionality questions (one for each committee member) followed by two Causation questions, or two Causation questions followed by two Intentionality questions. The order in which committee members appeared in the questions was randomized across participants but fixed within-participant. Participants were asked how much they agreed with the following statements, on a 1-7 likert scale (1: strongly disagree, 7: strongly agree):

- [Mr A /Mr B] intentionally gave Professor Smith new computers.

- The fact that [Mr A / Mr B] wanted Professor Smith to get new computers caused her to get new computers[14].

[14] For exploratory purposes, half the participants were asked the following causation question instead: "Professor Smith got new computers because Mr A wanted her to get new computers".

- The fact that [Mr A / Mr B] voted Yes tells us that it was important for him that Professor Smith get new computers.

Additionally, participants were asked the following two comprehension questions, on the same page just below the intentionality question:

- In order for a request to be approved, how many committee members need to vote Yes? (One / Two / Three / Impossible to tell)

- Both committee members usually reach the same decision most of the time. (True / False / Impossible to tell)

Participants who failed either question were excluded from analysis.

5.5.2) Results

For intentionality ratings, we found the predicted abnormal inflation effect in the conjunctive structure, as well as the predicted abnormal deflation effect in the disjunctive structure. The cross-over interaction was also statistically significant (see Figure 7).

In the conjunctive structure, intentionality ratings were higher for the norm-violating agent ($M = 5.67$, $SD = 1.67$) than for the norm-conforming agent ($M = 5.22$, $SD = 1.94$), $t(71) = 3.01$, $p = .004$, $d_z = .35$. By contrast, in the disjunctive structure, intentionality ratings were lower for the norm-violating agent ($M = 4.83$, $SD = 2.00$) than for the norm-conforming agent ($M = 5.39$, $SD = 1.61$), $t(58) = -2.41$, $p = .02$, $d_z = -.31$.

A 2*2 mixed Anova showed a significant interaction between Agent and Causal Structure on intentionality ratings, $F(1, 129) = 14.22$, $p < .001$, $\eta^2_{partial} = .10$.

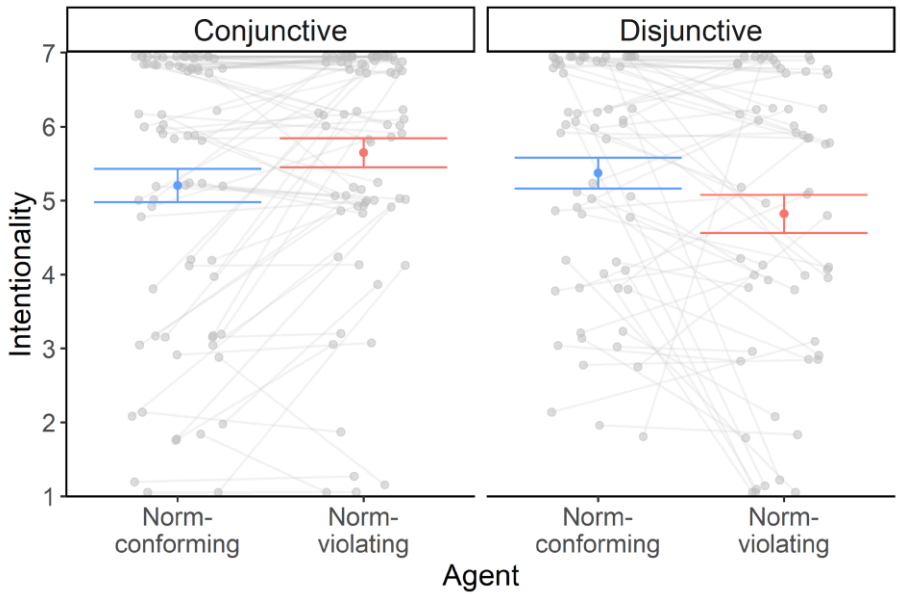

Figure 7: Intentionality ratings for norm-conforming (Mr A) and norm-violating (Mr B) agents, for conjunctive and disjunctive structure. Error bars represent standard errors of the mean. Individual data points are jittered for better visibility.

We also replicated the known pattern of results for causation judgments (see Figure 8). In the conjunctive causal structure, causation ratings were higher for the norm-violating agent ($M = 5.57$, $SD = 1.32$) than for the norm-conforming agent ($M = 4.58$, $SD = 1.53$), $t(71) = -5.90$, $p < .001$, $d_z = .69$. By contrast, in the disjunctive causal structure, causation ratings were lower for

the norm-violating agent (M = 4.20, SD = 1.88) than for the norm-conforming agent (M = 4.95, SD = 1.74), t(58) = -2.71, p = .009, d_z = -.35.

A 2*2 mixed Anova showed a significant interaction between Agent and Causal Structure on causation ratings, F(1,129) = 31.20, p <.001, $\eta^2_{partial}$ = .19.

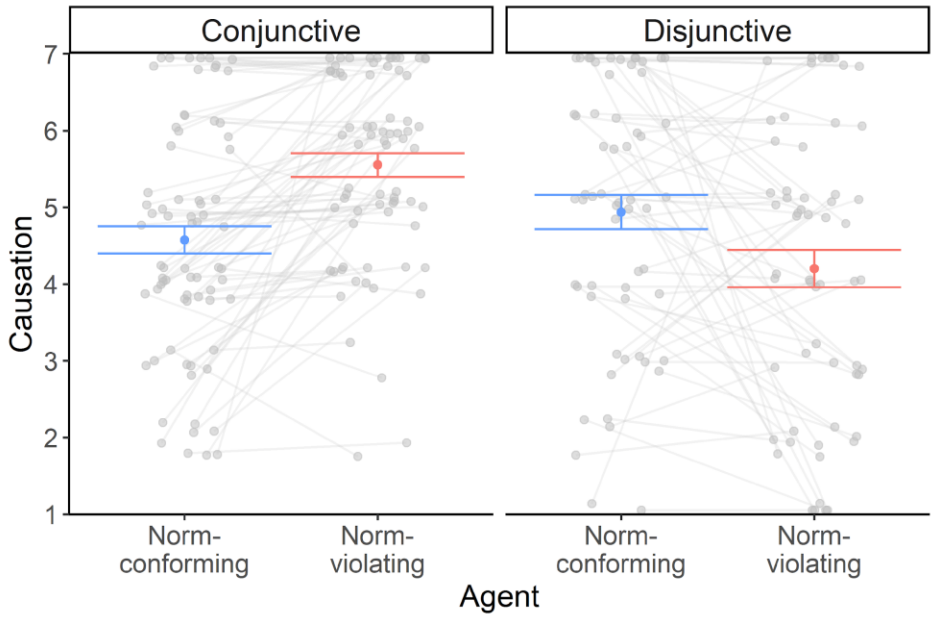

Figure 8: Causation ratings for norm-conforming (Mr A) and norm-violating (Mr B) agents, for conjunctive and disjunctive structure. Error bars represent standard errors of the mean. Individual data points are jittered for better visibility.

Given the similarity in the pattern of results between causation and intentionality ratings, it seems likely that intentionality judgments are shaped by normality considerations because normality influences causation judgments. But we wanted to test a possible alternative

interpretation for the cross-over interaction depicted in Figure 7. In the disjunctive structure, since Mr A is known to always vote Yes, and it takes only one committee member for a request to be approved, participants may reason that Mr B knows that he cannot change the outcome of the vote. Therefore, they may suspect that Mr B has little incentive to care, and has decided to vote randomly. If participants reason in this way, then they will think that in the disjunctive causal structure, Mr B's vote does not tell us much about how much he values Prof Smith's getting new computers. If intentionality judgments are driven by inferences about attitudes (Uttich & Lombrozo, 2010), then this line of reasoning will lead participants to give lower intentionality ratings to Mr B in the disjunctive structure compared to the conjunctive structure, which could explain our results.

The results for the inference ratings ("The fact that [Mr A / Mr B] voted Yes tells us that it was important for him that Professor Smith get new computers.") are not consistent with this interpretation (see Figure 9). Participants consistently judged that we learn more about the norm violator's attitude than about the other agent's attitude, and this effect was of the same size in both causal structures. In a 2*2 mixed Anova with Causal Structure and Agent as predictors, and inference ratings as outcome variable, there was a main effect of Agent, $F(1, 129) = 87.12$, $p < .001$, such that participants gave higher inference ratings for the norm-violating agents ($M = 6.07$, $SD = 1.27$) compared to the norm-conforming agent ($M = 4.49$, $SD = 1.80$). There was also a main effect of Causal Structure, $F(1, 129) = 5.09$, $p = .03$, such that participants gave higher inference ratings in the conjunctive ($M = 5.49$, $SD = 1.59$) than the disjunctive causal structure ($M = 5.02$, $SD = 1.89$). However, there was no interaction between Agent and Causal Structure, $F(1,129) = 0.95$, $p = .33$.

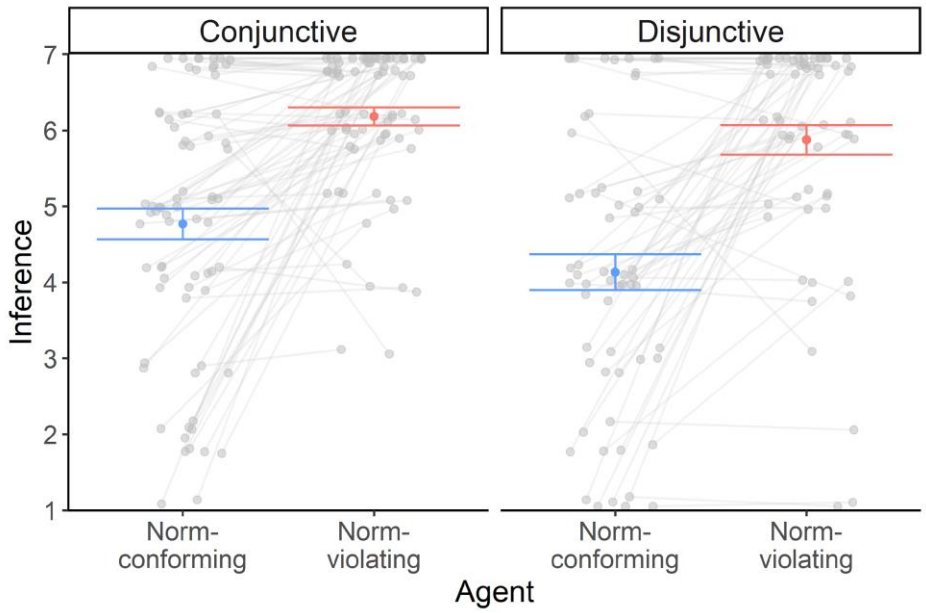

Figure 9: Inference ratings for norm-conforming (Mr A) and norm-violating (Mr B) agents, for conjunctive and disjunctive structure. Error bars represent standard errors of the mean. Individual data points are jittered for better visibility.

We also conducted a series of Anovas to check for order and wording effects. Most tests were negative, with the following exceptions. There was a 2x2 interaction between the Wording of the causation question and the identity of the Agent on the causation ratings, $F(1, 127) = 6.33$, $p = .01$; such that, averaging across causal structures, the norm-violating agent was rated as more causal than the norm-conforming agent, but only for the "Caused" wording of the causation question. There was also a 2x2x2 interaction between the Order of presentation of agents (Mr A vs Mr B first), the Causal Structure, and the identity of the Agent, on the causation ratings, $F(1,127) = 4.60$, $p = .03$; such that in the disjunctive causal structure, the norm-violating agent

was seen as less causal than the norm-conforming agent only when the norm-violating agent was presented last; see Supplementary Information (https://osf.io/dp5xr/?view_only=64e21726c536419c9942ac2ca1aca9c1) for details.

5.5.4) Study 5B

According to our account, intentionality judgments are computed from a mental representation of the strength with which the agent's attitude toward the outcome caused the outcome. The abnormal inflation and abnormal deflation effects we found for intentionality judgments suggest that, when answering the intentionality question, people computed how much the agent's attitude toward the outcome caused the outcome.

But maybe people were engaging in another kind of causal strength computation, in which mentalizing plays no role. Our intentionality question asked participants how much they agreed that the agent intentionally gave computers to Professor Smith; therefore, some of the variation in intentionality ratings might be due to variation in agreement that the agent *gave* computers to Professor Smith. Furthermore, people may have been interpreting "the agent gave computers to Professor Smith" as "the fact that the agent voted Yes caused Professor Smith to get new computers", and where "the fact that the agent voted Yes" was construed in a purely mechanical sense. Under these two assumptions, one would predict abnormal inflation and abnormal deflation effects for the intentionality question, even if people were not computing the causal dependence between the outcome and the agent's attitude toward the outcome.

This interpretation is *a priori* unlikely, because researchers working on the semantics of "giving" have suggested that intentionality is itself an important part of the concept (Newman, 1996). Nonetheless, we wanted to make sure that in the scenario we used in study 5, people did

not interpret “the agent gave computers to Professor Smith” as meaning simply “the fact that the agent voted Yes caused Professor Smith to get new computers”.

To that end, we conducted study 5B (see supplementary information at https://osf.io/42x7h/?view_only=64e21726c536419c9942ac2ca1aca9c1), where we used a similar scenario as in study 5, except that the committee members could not know that their vote had the potential to cause a professor to get funding. We asked half the participants to rate whether the vote of each committee member *caused* Professor Smith to get new computers, and the other half to rate whether the committee member *gave* new computers to Professor Smith. We found that while participants gave high causation ratings (their mean is above 5 on a 1-7 scale), they were reluctant to agree that the committee member gave computers to Professor Smith (mean ratings below 3). Additionally, we observed an abnormal inflation effect for causation ratings, but not for ratings of giving. These results suggest that participants in study 5 (which used almost the same scenario as study 5B) did not interpret “the agent gave computers to Professor Smith” as simply meaning “the fact that the agent voted Yes caused Professor Smith to get new computers”.

Therefore, the specific pattern of effects of normality on intentionality ratings found in study 5 probably indicates that people computed how much the agent’s attitude toward the outcome caused the outcome.

5.5.5) Discussion

Results of Study 5 provide evidence in favor of a novel prediction of our account: the influence of normality on intentionality judgments depends on the causal structure of the situation. In a conjunctive structure, we found that participants judged an agent violating a statistical norm as more intentional than a norm-conforming agent. In a disjunctive causal structure, this effect was reversed: the norm-conforming agent was judged as more intentional than the norm-violating agent. These results are difficult to explain on any account that predicts a general bias to consider norm violations as more intentional (e.g. Alicke & Rose, 2010; Hindricks, 2014; Uttich & Lombrozo, 2010).

5.6) Study 6: Intentionality requires a domain-specific causal pathway

In the studies described so far, judgments of intentionality closely track judgments of causation. As a reviewer observed, our framework also predicts that intentionality and causation judgments will sometimes diverge from each other. Specifically, if an agent's attitude toward an outcome causes that outcome in a way that deviates from the causal model implicit in commonsense psychology, people will tend to deny that the outcome was brought about intentionally, even when they judge that it was caused by the agent's attitude toward that outcome.

We test this prediction in study 6. We designed a vignette about a futuristic corporation whose employees have brain implants. The AC in the building is regulated as a function of how the employees feel about the temperature, as recorded by their brain implants. In one version of the vignette, the causal link between the agents' attitude and the outcome conforms to the causal model in commonsense psychology (the employee needs to make a decision in order for the AC to be turned on). In the other vignette, the causal link is deviant, in the same way as is the causal link in the **King** example from section 3.1. Specifically, the brain implant automatically detects and implements the employee's desires.

We predict that in the normal link condition, judgments of whether an agent intentionally turned on the AC will be relatively close to judgments of whether the AC was turned on because the agent wanted it to. By contrast, in the deviant link condition, judgments of intentionality will be lower than judgments of causation. In addition, we also manipulated the normality of the agents' attitudes, in an attempt to replicate the abnormal deflation effect found in study 5.

5.6.1) Methods

Participants

Our main prediction is an interaction between Question Type (causation vs intentionality, within-subject) and causal link (normal vs deviant, between-subject). Although we expected a large effect, we conservatively set our intended sample size (228 participants) so that we would be able to detect a small effect ($\eta^2_{partial}$ = .02) with 99% power. Anticipating an exclusion rate similar to study 5 (33%), we set our recruitment target at 350 participants.

We recruited 349 US residents from Mechanical Turk. We excluded from analysis 86 participants failing either a catch item (N=3), or either of three comprehension questions (N=18, 29, 53), leaving a final sample of 263 participants (139 female, 1 unspecified).

Design

We manipulated Causal Link between-subjects. Half the participants read a story in which the computer can automatically detect employee's attitudes toward making the room cooler (*deviant link condition*). The other half of participants read a story in which employees need to formulate a request in their head in order to communicate their desire to make the room cooler (*normal link condition*).

We also manipulated the normality of agents' attitudes, in a similar manner as in study 5. All participants read about two employees, Mr A and Mr B. Mr A was described as very sensitive to heat: thus, his desire to make the room cooler is statistically normal. By contrast, Mr B was described as usually indifferent to the room temperature, thus his current desire to make the room cooler is statistically abnormal.

The story we used describes a disjunctive causal structure: the computer activates the AC if it detects that at least one agent would like to make the room cooler. In this kind of setting, counterfactual models of causal judgment predict that the normal agent (here, Mr A) should be viewed as more causal than the abnormal agent (Mr B).

For each agent, participants were asked a Causation question (probing whether the agent's attitude caused the room to be cooler), and an Intentionality question (probing whether the agent intentionally made the room cooler).

Stimuli and Procedure.

Participants read the following vignette:

"In the FutureCorp building, employees have a chip inside their head that can read their brain activity. [On the basis of your brain activity, the chip can predict how you would respond to the question "would you like the room to be cooler?" if someone were to ask you / If you feel that it is too hot, you can request the AC to be turned on by thinking in your head "I would like the room to be cooler"]. Whenever the chip detects that you would like the room to be cooler, it sends this information to the computer that regulates the room's temperature. Because the chip is perfectly accurate, [the computer can determine whether you want the room to be cooler without actually asking you / the computer can determine whether you want the room to be cooler if you formulate the request in your head]. Employees know how the system works.

For every room in the building, the rule is that the computer turns on the AC in that room if it detects that at least one person in the room would like the room to be cooler.

In room 42, there are two employees, Mr A and Mr B. While Mr A is very sensitive to heat, Mr B usually doesn't really care. Today is a very hot day, and the system accurately detects that both employees [would be in favor of turning on the AC if they were asked / made a request in their head to turn on the AC]. Since at least one employee in the room would like the room to be cooler, the system turns on the AC, and the room gets cooler."

Participants were then asked two Causation questions (one for each employee) and two Intentionality questions, on separate pages. Half the participants saw the two Causation questions first, followed by the two Intentionality questions; this order was reversed for the other half. The order in which the employees appeared in the questions was randomized across participants but

fixed within-participant. Participants were asked how much they agreed with the following statements, on a 1-7 likert scale (1: strongly disagree, 7: strongly agree):

-[Mr A / Mr B] intentionally made the room cooler

-The room got cooler because [Mr A/ Mr B] wanted the room to be cooler

Additionally, participants were asked the following three comprehension questions, on the same page just below the first question:

-In order for the AC to be activated, how many people need to be in favor of making the room cooler? (0, 1, 2, 3)

-Both employees are equally sensitive to heat (True / False / Impossible to tell)

-The computer cannot automatically detect what employees want: they need to formulate a request in their head (True / False)

Participants who failed any of these questions were excluded from analysis.

5.6.2) Results

We first look at whether our main prediction is supported. Then we turn to the effect of normality.

In the normal link condition, participants' mean causation and intentionality ratings were almost identical. By contrast, and as predicted, in the deviant link condition intentionality ratings were much lower than causation ratings (see figure 10).

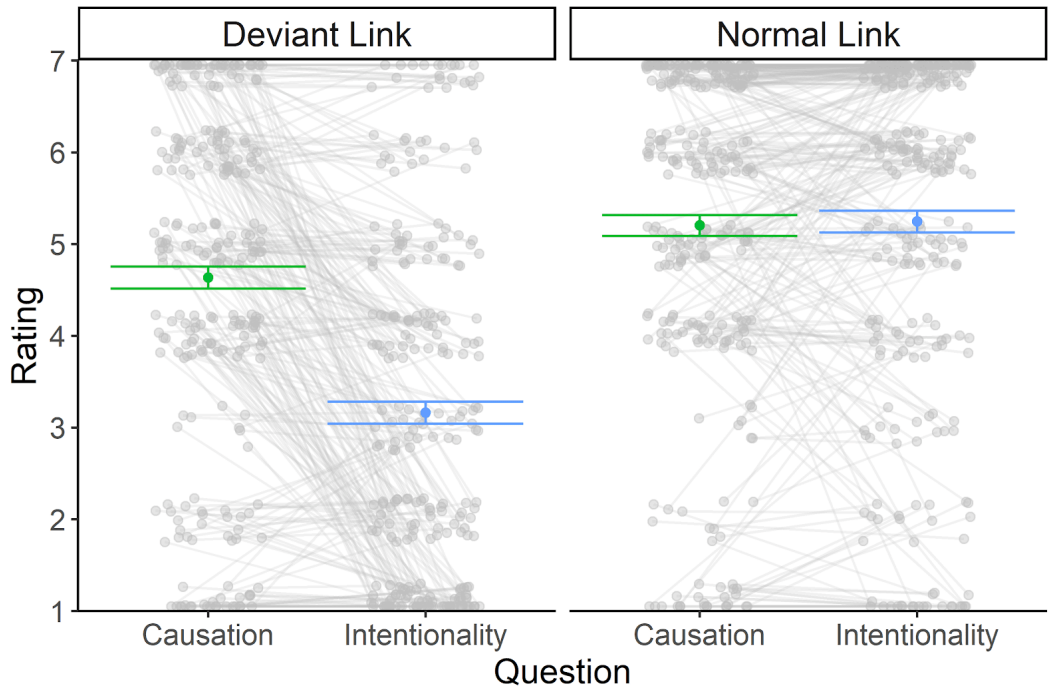

Figure 10. Causation and intentionality ratings for the deviant and normal causal links. Colored points represent average judgment, averaging across agents (e.g. a green point represents the average causation rating for Mr A and Mr B combined). Error bars represent standard errors of the mean. Individual data points are jittered for better visibility.

We analyzed the data with a 2 (Causal link) * 2 (Question) * 2 (Agent) mixed Anova.

There was a main effect of Causal link, $F(1, 261) = 79.9$, $p < .001$, such that participants in the Normal link condition gave higher ratings ($M = 5.24$, $SD = 1.92$) than participants in the Deviant link condition ($M = 3.90$, $SD = 1.92$). There was also a main effect of Question, $F(1, 783) = 58.6$, $p < .001$, such that participants gave higher ratings of causation ($M = 4.93$, $SD = 1.93$) than intentionality ($M = 4.23$, $SD = 2.22$). Crucially, this effect was much higher in the Deviant link condition (causation: $M = 4.64$, intentionality: $M = 3.15$) than in the Normal link

condition (causation: M = 5.22, intentionality: M = 5.26), as shown by an interaction between Causal link and Question, $F(1, 783) = 68.8$, $p < .001$, $\eta^2_{partial} = .081$.

We also conceptually replicate the abnormal deflation effect found in study 5. There was a main effect of Agent, $F(1, 783) = 261.7$, $p < .001$, such that participants gave higher ratings for the norm-conforming agent, Mr A (M = 5.33, SD = 1.77) than the norm-violating agent, Mr B (M = 3.83, SD = 2.15). This abnormal deflation effect held for both intentionality and causation ratings. Intentionality ratings for the norm-conforming agent, Mr A (M = 4.90, SD = 2.05) were higher than for the norm-violating agent, Mr B (M = 3.55, SD = 2.17), $t(262) = 10.72$, $p < .001$, $dz = .66$. The norm-conforming agent was also judged as more causal (M = 5.75, SD = 1.31) than the norm-violating agent (M = 4.11, SD = 2.10), $t(262) = 11.6$, $p < .001$, $dz = .72$. Interestingly, abnormal deflation was equally strong in the deviant link and the normal link conditions, for both intentionality and causation ratings: there was no significant interaction between Causal link and Agent on intentionality ratings, $F(1, 261) = .287$, $p = .593$, or on causation ratings, $F(1, 261) = 2.235$, $p = .136$.

We also ran an Anova with Causal link, Agent, Question type, Order of questions (causation or intentionality first), and Order of agents (Mr A or Mr B first) as predictors, and found no significant interactions involving the order in which the questions were asked.

In an exploratory analysis, we also found that the participants giving the highest causation ratings also tended to give the highest intentionality ratings, but that this correlation was higher in the normal compared to the deviant link condition. Across participants, the

correlation between intentionality and causation judgments for Mr A was descriptively higher[15] in the normal link ($r(132) = .63$) than in the deviant link condition ($r(127) = .44$), $p = .06$. The correlation between intentionality and causation judgments for Mr B was significantly higher in the normal link ($r(132) = .74$) than in the deviant link condition ($r(127) = .47$), $p = .005$.

5.6.3) Discussion

We find that judgments of intentionality diverge from judgments of causation when the causal link between an agent's attitude and an outcome goes outside of the typical causal model of commonsense psychology. In a situation where a machine can automatically detect and implement an agent's desire, participants were much more willing to agree that the agent's attitude toward the outcome caused the outcome, than to agree that the agent acted intentionally.

On the other hand, there was one way that judgments of intentionality were similar to judgments of causation. Participants tended to attribute lower intentionality to the norm-violating agent than the norm-conforming agent, even when the causal link between attitude and outcome was deviant. This pattern reflected the one found for causation judgments. One possible interpretation is that deviance from the causal model of commonsense psychology attenuates intentionality judgments, but does not completely disconnect them from computations of causal strength. We leave it to future research to shed more light on this finding.

[15] To compare the strength of the two correlations, we first z-scored the intentionality and causation scores within each Causal link condition, and then computed the interaction between causation scores and causal link in a multiple regression predicting intentionality scores. The p-value we report is the p-value for that interaction.

6) General discussion

What does it mean to do something *intentionally*? We have argued that, to the human mind, the concept is defined in the following way: "an agent did X intentionally if the agent's attitude toward X caused X, and caused X according to the typical causal model implicit in our commonsense psychology*"*

Philosophers and cognitive scientists have found it difficult to define intentional action, because people's intuitions about the use of the concept are very complex, such that it is easy to find counterexamples to a given definition. We suggested that this complex pattern of intuitions can be explained by taking a closer look at the building blocks of the concept.

According to our proposal, these building blocks include the causal model of commonsense psychology, and the intuitive concept of causation. By incorporating recent insights from cognitive science models of these building blocks, our account can (i) provide a unifying explanation for a variety of known patterns in human intuitions about intentional action, such as the side-effect effect; and (ii) make successful novel predictions.

Below we discuss limitations of our account, address alternative explanations for our data, and explore implications for the learnability of Theory of Mind concepts.

6.1) Limitations

On our account, a full understanding of the folk concept of intentional action will have to rely on a full understanding of its building blocks, such as mechanisms for causal judgment.

Current models of causal cognition have great explanatory and predictive power, but are still incomplete; as a result, one may not currently be able to provide a fully mechanistic explanation for every intuition that people have.

Consider for instance a case where Jake shoots at his aunt from a great distance with the intention of killing her, and manages to reach his target by sheer luck, despite being a poor marksman. Most people intuitively judge that Jake intentionally killed his aunt (Knobe, 2003b). This result makes sense in our framework, since intuitively the aunt died *because* Jake wanted to kill her. Yet to have a fully mechanistic explanation, we would need to explain where this causal intuition comes from. Computational models of causal judgment do not make a clear prediction in this case. On the one hand, Jake's murderous desire is clearly morally abnormal, so we should expect that people will judge it to be causal (because people will readily generate counterfactuals where Jake does not have the desire to kill his aunt, and in these counterfactuals his aunt does not dies). On the other hand, the shot reached Jake's aunt by a stroke of luck: it is easy to entertain counterfactuals where Jake shoots and misses his target; this will make people less likely to think of his murderous desire as causal.

Therefore we have two variables (the agent's abnormal morality, and his lack of control over the outcome), which theoretically should pull causation judgments in opposite directions, but current models of causal judgment do not make clear predictions about which variable will have a stronger effect. In order to explain why it feels intuitive to consider Jake's murderous desire as a strong cause of his aunt's death, we must make ad-hoc assumptions, for instance, the assumption that in this case people are more likely to generate counterfactuals as a function of moral normality. Hopefully, as cognitive scientists develop increasingly more accurate models of causal cognition, our account will be able to make more fine-grained, mechanistic predictions.

6.2) Alternative explanations

Many of the empirical tests we have presented here involved showing that intentionality judgments and causation judgments track each other. Our interpretation of these results is that intentionality judgments are computed from a mental representation of the causal dependence between an outcome and the agent's attitude toward this outcome. Are there plausible alternative explanations?

Maybe the causal arrow runs in the reverse direction: causal judgments are computed from a mental representation of intentionality. For instance, our questions about causation may have sounded unnatural and confusing, and therefore participants defaulted to interpreting them as questions about intentionality.

Maybe a third variable explains why causation and intentionality judgments track each other. Notably, researchers have found that judgments about intentional action and judgments about causation are both influenced by the degree to which people regard certain counterfactuals as relevant when asked to consider how things could have gone differently (Phillips, Luguri & Knobe, 2015). This effect of counterfactual relevance seems quite wide-ranging, since Phillips et al. (2015) also found that it has an impact on people's judgments of freedom and judgments about the doing/allowing distinction. Therefore, one natural interpretation is that intentionality and causation judgments track each other simply because they are both independently impacted by judgments of counterfactual relevance.

We see at least three reasons to favor our "causation first" account over these two alternatives. First, even if our minimalist account turned out to be incomplete, it is almost

undeniable that causation has to be a central component of intentional action. We challenge the reader to find an example of a case where an agent intentionally does X, yet there is absolutely no causal connection between X's attitude toward X and X happening. By contrast, it is equally obvious that causation does not *require* intentionality – people spontaneously attribute causation in situations devoid of any mental states, such as physical collisions between billiard balls, and the mechanism for this inference appears to be present in infants as young as 6 months of age (Michotte, 1963; Leslie & Keeble, 1987; Gerstenberg & Icard, 2019).

Second, the results of study 5 are difficult to interpret on the alternative explanations sketched above. On an 'intentionality first' account, there are no *a priori* reasons why normality considerations would interact with causal structure in shaping intentionality judgments (to our knowledge, no existing account of intentional action predicts this interaction effect). On a "third variable" account involving counterfactual relevance, there is no *a priori* reason to expect the interaction effect either. Indeed, preliminary evidence suggests that people make the same judgments of counterfactual relevance in disjunctive and conjunctive cases: in both causal structures, they tend to view the norm-violating event as more relevant (Kominsky & Phillips, 2019, experiment 2). Therefore, in order to explain the interaction effect, we must appeal to a richer explanatory framework. Computational models of causal judgments such as Icard et al. (2017) and Quillien (2020) provide exactly such a framework: they predict such an interaction effect *despite* assuming that people generate counterfactuals in the same way in both kinds of causal structure.

Third, the results of study 6 show that causation and intentionality judgments do not always track each other. We constructed a situation where an agent's attitude causes an outcome in a way that deviates from the causal model of commonsense psychology, but otherwise

satisfies the intuitive, domain-general concept of causation. As predicted by our account, people were reluctant to judge that the agent intentionally brought about the outcome, but tended to agree that the agent's attitude caused the outcome. This pattern of intuitions would be unlikely if causation judgments were computed from intentionality judgments.

Therefore, 'causation first' is the account that best explains our data: intentionality judgments track causation judgments because they are computed from a mental representation of causation.

6.3) Minimality and Learnability

Intuitively, a strong appeal of our account is its minimality. Having said that, we are not arguing that it would be simple to teach the meaning of "intentionally" to a machine. Indeed, our account predicts exactly the contrary: it would be very difficult to teach the concept to an artificial intelligence with a "blank-slate" architecture. Our account is very simple because it assumes that the concept of intentional action is constructed in a relatively straightforward way from pre-existing building blocks. These building blocks include complex cognitive mechanisms for reasoning about the minds of others and for making causal attributions.

In other words, we do not aim to provide a comprehensive and transparent definition of the folk concept of intentional action, that could be hard-wired into a computer devoid of any other specialized knowledge. Rather, we are trying to decipher the recipe by which the reliably-developing human brain acquires the concept. Thus, one appealing feature of our theory is that it suggests a solution to a learnability problem (Jackendoff, 1989; Pinker, 1989) that we think has been neglected in existing debates about the meaning of the concept. Existing accounts of

"intentionally" can be quite complex, for instance requiring at least 5 necessary components for something to intentional (e.g. Malle & Knobe, 1997), or positing that people attribute different meanings to the word, with the relevant meaning being determined by the context (e.g. Cova et al., 2012; Cushman & Mele, 2008). How do children manage to acquire such a complex concept from the linguistic stimuli they are exposed to? And why do they acquire this very concept (or set of concepts) as opposed to any other? As far as we know, most existing theories are silent about these questions.

By contrast, if, as we suggest, the folk concept of intentional action is built in a relatively simple way from a set of pre-existing building blocks, then we can start to sketch an account of how children acquire the concept. Here is one suggestion. Assume that children already possess a set of Theory of Mind mechanisms, an intuitive concept of causation, and that they are predisposed to infer that, within the psychological domain, words that refer to a link between two entities refer to a *causal* link[16]. Therefore, when they come to understand that intentionality refers to some kind of relationship between an agent's attitude toward X and the occurrence of X, they spontaneously assume that an agent does X intentionally if his attitude toward X caused X.

7) Conclusion

People can use the word "intentionally" in very strange ways. Our intuitions about whether something is intentional are swayed by moral considerations, are pulled one way or

[16] A predisposition to assume that mentalizing concepts involve causal links also makes sense of the fact that concepts such as 'perceiving', 'remembering' and 'knowing' seem to have a causal component (see Grice, 1961; Martin & Deutscher, 1966; Goldman, 1967. For instance, cross-cultural evidence suggests that people everywhere think that justified true belief in p that is not caused by p does not constitute knowledge, Machery et al., 2017).

another depending on the amount of control an agent exerts, and are influenced by how circuitous the causal chain between the agent and the outcome is. Intentionality requires a relevant belief, but the latter can be present in very small doses. Norm-violating actions are judged as *more* intentional than norm-conforming actions -- except when they are judged as *less* intentional.

These seemingly erratic intuitions can be anxiety-inducing. One might conclude that our commonsense psychology is fundamentally moralistic; that linguistic meaning is hopelessly entangled in its context; or that motivational and pragmatic factors constantly warp our intuitions about the proper extension of words.

We think such anxiety might be misplaced. Instead, we view the strangeness of "intentionally" as emerging naturally from the core structure of the concept. The way people use the concept of intentional action offers a fascinating window on some of the building blocks that make up human thought: it lets us glimpse into our implicit causal model of the mind, and the algorithms with which we assign causes to events.

Appendix

A.1) Statements used in Study 1

Anne was sweating

Anne was yawning during the lecture

Anne was grinding her teeth during the test

Anne had a craving for cherries after dinner

Anne believed that she had the flu

Anne was in a great mood today

Anne was infatuated with Ben

Anne was worrying about the test results

Anne got admitted to Princeton

Anne interrupted her mother

Anne ignored Greg's arguments

Anne drove way above the speed limit

Anne applauded the musicians

Anne greeted her uncle politely

Anne refused the salesman's offer

Anne stole a pound of peaches

Anne asked Mike out for dinner

Anne invited Sue to have lunch with her

Anne watered her new plants

A.2) Additional vignettes used in Study 3

Shooter. Bob wants to kill Alice. After weeks of careful study of her daily routine, he finally has her in the line of sight of his rifle while she is walking in a quiet area of town. As he steadies his aim, he suddenly realizes that, in his excitement at the thought of carrying out his plan, he forgot to put bullets in his rifle before leaving his house that morning. Although he is now convinced that shooting will have no effect, he decides to pull the trigger anyway to release his anger.

Unbeknownst to him, a few bullets had been actually left in the rifle from one of his training sessions a few days before. His shot is perfectly accurate, and sends a bullet right through Alice's heart. To Bob's surprise, Alice dies instantly.

Belief question: "By pulling the trigger, Bob believed that he would kill Alice."

Intentionality question: "By pulling the trigger, Bob intentionally killed Alice"

Causation question: "Alice died because Bob wanted to kill her"

Comprehension question: "The bullets in the rifle were put there in the morning of the event" (True/False/Impossible to tell)

Sabotage. John is a worker at a power plant. In the control room of the power plant, there is a red button that engineers can push to cause the main reactor to shut down. As an extra safety feature, in addition to pressing the red button, shutting the reactor also requires a special key card. One day, John sees that there is nobody in the control room. He hates his boss and wants to make him look incompetent, so he decides to sneak in and press the red button.

He is aware that he doesn't have the special key card, and that therefore pressing the red button will not trigger the reactor shutdown. Yet John really wants to shut down the main reactor, so he presses the red button anyway.

As it happens, because of an oversight on the part of the engineers, the safety feature was not yet in place. To John's surprise, the main reactor shuts down, creating considerable confusion.

Belief question: "John believed that he would shut down the main reactor"

Intentionality question: "John intentionally shut down the main reactor"

Causation question: "The reactor shut down because John wanted to shut down the reactor"

Comprehension question: "There were a few engineers in the control room" (True / False / Impossible to tell)

References

Adams, F., & Steadman, A. (2004). Intentional action in ordinary language: Core concept or pragmatic understanding? *Analysis*, 64, 173-181

Alicke, M. (2008). Blaming badly. *Journal of Cognition and Culture*, *8*(1-2), 179-186

Alicke, M., & Rose, D. (2010). Culpable control or moral concepts?. *Behavioral and brain sciences*, 33(4), 330.

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *The Journal of Philosophy*, *108*(12), 670-696.

Anscombe, G. E. M. (1957). *Intention*. Harvard University Press.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states?. *Psychological review*, 116(4), 953.

Aristotle (330BC / 2009). *The Nicomachean ethics*. Oxford: Oxford University Press.

Bennett, D. (1965). Action, Reason and Purpose. *Journal of Philosophy, 62*, 85-95.

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

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Brand, M. (1984). *Intending and acting: Toward a naturalized action theory.* Cambridge, MA: MIT Press.

Bratman, M. E. (1987). *Intention, plans, and practical reason.* Cambridge, MA: Harvard Univ. Press.

Bratman, M. (1984). Two faces of intention. *The Philosophical Review*, *93*(3), 375-405.

Baillargeon, R., Scott, R. M., & Bian, L. (2016). Psychological reasoning in infancy. *Annual review of psychology*, *67*, 159-186.

Bear, A., Bensinger, S., Jara-Ettinger, J., Knobe, J., & Cushman, F. (2020). What comes to mind? *Cognition*, 194 , 104057.

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition*, 167 , 25-37.

Beebe, J. R., & Buckwalter, W. (2010). The epistemic side-effect effect. *Mind & Language*, *25*(4), 474-498.

Byrne, R. M. (2016). Counterfactual thought. *Annual review of psychology*, *67*, 135-157.

Cova, F., Dupoux, E., & Jacob, P. (2012). On doing things intentionally. *Mind & Language*, *27*(4), 378-409.

Cova, F. (2016). The folk concept of intentional action: Empirical approaches. In W. Buckwalter & J. Sytsma (Eds.), *Blackwell Companion to Experimental Philosophy*. Wiley-Blackwell.

Cosmides, L. (1985). *Deduction or Darwinian algorithms? An explanation of the "elusive" content effect on the Wason selection task*. Unpublished doctoral dissertation, Harvard University

Cushman, F. (2015). Deconstructing intent to reconstruct morality. *Current Opinion in Psychology*, *6*, 97-103.

Cushman, F., & Mele, A. (2008). Intentional Action: two-and-a-half concepts?. *Experimental philosophy*, *171*.

Davidson, D. (1963). Actions, reasons, and causes. *The journal of philosophy*, 60(23), 685-700.

Davidson, D. (1980). *Essays on Actions and Events*. Oxford University Press.

Dennett, D. C. (1987). *The intentional stance*. MIT press.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in cognitive sciences, 7*(7), 287-292.

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

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. *Oxford handbook of causal reasoning*, 515-548.

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological science, 28*(12), 1731-1744.

Gerstenberg, T., & Icard, T. (2019). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General.*

Goldman, A. I. (1967). A causal theory of knowing. *The Journal of Philosophy*, *64*(12), 357-372.

Goldman, A. I. (1970). *Theory of human action*. Princeton University Press.

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

Grice, H. P. (1961). The causal theory of perception. *Proceedings of the Aristotelian Society*, 35, 121-168.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, and L. A. Paul (Eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

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

Halpern, J. (2016). *Actual causality*. MIT Press.

Hart, H. L. A., & Honoré, T. (1985). *Causation in the Law*. OUP Oxford.

Heider, F. (1958). The psychology of interpersonal relations. Psychology Press.

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

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

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, *98*(6), 273-299.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, 106 (11), 587-612.

Hindriks, F. (2014). Normativity in action: How to explain the Knobe Effect and its relatives. *Mind & Language*, 29, 51-72.

Hume, D. (1740 / 1978). Treatise of human nature (L. A. Selby-Bigge, Ed.; 2nd ed.). Oxford University Press.

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

Jackendoff, R. (1989). What is a Concept, that a Person May Grasp It?. *Mind & Language*, *4*(1-2), 68-102.

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.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, 123, 101334.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 371–388). Hillsdale, NJ: Erlbaum.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological review*, 93 (2), 136.

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis*, *63*(279), 190-194.

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical psychology*, *16*(2), 309-324.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*(4), 315-329.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, 137 , 196-209.

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.

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

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality?. *Cognition*, *25*(3), 265-288.

Leslie, A. M. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. *Mapping the mind: Domain specificity in cognition and culture*, 119-148.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in 'theory of mind'. *Trends in cognitive sciences*, *8*(12), 528-533.

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.

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.

Livengood, J., & Machery, E. (2007). The folk probably don't think what you think they think: Experiments on causation by absence. *Midwest Studies in Philosophy, 31*, 107-127.

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

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., ... & Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PLoS one*, *9*(3), e92160.

Machery, E. (2008): The folk concept of intentional action: philosophical and experimental issues. *Mind & Language*, 23, 165–189.

Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., ... & Hashimoto, T. (2017). Gettier Across Cultures. *Noûs*, *51*(3), 645-664.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*(2), 101-121.

Martin, C.B., & Deutscher, M. (1966). Remembering. *Philosophical Review*, 75(2): 161–96.

Mele, A. (2001). Acting intentionally: probing folk notions. In *Intentions and Intentionality: Foundations of Social Cognition*, ed. B. Malle, L. Moses and D. Baldwin, 27–43. Cambridge, MA: MIT/Bradford

Mele, A. R., & Cushman, F. (2007). Intentional action, folk judgments, and stories: Sorting things out. *Midwest Studies in Philosophy*, *31*, 184-201.

Mele, A. (2009). Causation, Action, and Free Will. In Beebee, H., Hitchcock, C., & Menzies, P. (Eds.). *The Oxford handbook of causation.* Oxford University Press.

Michotte, A. (1963). *The perception of causality*. Basic books.

Mill, J. S. (1856). *A System of Logic, Ratiocinative and Inductive*. London: John W. Parker and Son.

Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PLoS one*, *14*(8).

Nadelhoffer, T. (2006). Bad acts, blameworthy agents and intentional actions: Some problems for jury impartiality. *Philosophical Explorations*, 9, 203-220.

Newman, J. (1996). *Give: A cognitive linguistic study*. Walter de Gruyter.

Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The Knobe effect revisited. *Mind & Language, 22*(4), 346-365.

Ossorio, P. G., & Davis, K. E. (1968). The self, intentionality, and reactions to evaluations of the self. In C. Gordon & K. J. Gergen (Eds.), *The self in social interaction.* New York:Wiley.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & Language, 24*, 586-604. doi:10.1111/j.1468- 0017.2009.01375.x

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30-42.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in cognitive sciences*.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. MIT press.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. Penguin.

Quillien, T. (2020). When do we think that X caused Y?. *Cognition. 205*

Quillien, T., & Barlev, M. (2021). Causal judgment in the wild: evidence from the 2020 US presidential election. *PsyArXiv*. https://doi.org/10.31234/osf.io/7w9re

Rosseel Y (2012). "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, **48**(2), 1–36.

Savage, L.J. (1954). *The foundations of statistics.* Wiley, New York.

Scott, R. M., & Baillargeon, R. (2013). Do infants really expect agents to act efficiently? A critical test of the rationality principle. *Psychological science*, *24*(4), 466-474.

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind.* Cambridge, Cambridge Univ. Press.

Sell, A., Sznycer, D., Al-Shawaf, L., Lim, J., Krauss, A., Feldman, A., ... & Tooby, J. (2017). The grammar of anger: Mapping the computational architecture of a recalibrational emotion. *Cognition*, *168*, 110-128.

Shaver, K. G. (1985). *The attribution of blame.* New York: Springer-Verlag.

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

Sousa, P., & Holbrook, C. (2010). Folk concepts of intentional action in the contexts of amoral and immoral luck. *Review of Philosophy and Psychology, 1*(3), 351-370.

Sousa, P., Holbrook, C., & Swiney, L. (2015). Moral asymmetries in judgments of agency withstand ludicrous causal deviance. *Frontiers in psychology, 6*, 1380.

Sripada, C. (2012). Mental states attribution and the side-effect effect. *Journal of Experimental Social Psychology, 48*, 232- 238. doi:10.1016/j.jesp.2011.07.008

Strickland, B. (2017). Language reflects "core" cognition: A new theory about the origin of cross-linguistic regularities. *Cognitive science*, *41*(1), 70-101.

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 Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(4), 814-820.

Tesser, A., Gatewood, R., & Driver, M. (1968). Some determinants of gratitude. *Journal of personality and social psychology*, *9*(3), 233.

Thalberg, I. (1984). Do our intentions cause our intentional actions? *American Philosophical Quarterly, 21*, 249–260.

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116*(1), 87-100.

Wertz, A. E., & German, T. C. (2007). Belief–desire reasoning in the explanation of behavior: Do actions speak louder than words?. *Cognition*, *105*(1), 184-194.

Weslake, B. (2015). A partial theory of actual causation. *The British journal for the philosophy of science.*

Wittgenstein, L. (1953). *Philosophical investigations*. John Wiley & Sons.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1-34.

Wright, J., & Bengson, J. (2009). Asymmetries in folk judgments of responsibility and intentional action. *Mind & Language*, 24, 237-251.