It has been said that "history is the story of events, with praise or blame" (Cotton Mather). But each of us tells the story of our own life in this way. We are evaluative creatures. We don't just simply see *what* people do. We can't help but think about *why* they did it, and whether it was a good idea. And when someone's actions affect our welfare, we tend to hold them responsible.

I study how people hold others responsible, how these judgments are grounded in causal representations of the world, and supported by counterfactual simulations. The links between responsibility, causation, and counterfactuals are deep. While judgments of credit or blame often come to us easily (and sometimes too quickly), they are founded on sophisticated inferences operating over a rich common-sense understanding of the world. To hold others responsible we have to 1) determine what causal role their action played in bringing about the outcome, and 2) infer what the action revealed about the person, such as their abilities, beliefs, and intentions. To answer the first question, we need a model of how the world works. To answer the second question, we need a model of how people work – an intuitive theory of how others make decisions which allows us to reason backward from observed actions to the underlying mental states that caused them (Gershman*, Gerstenberg*, Baker, & Cushman, 2016; Gerstenberg & Tenenbaum, in press). In my research, I formalize people's mental models as computational models that yield quantitative predictions about a wide range of situations. To test these predictions, I use a combination of large-scale online experiments, interactive experiments in the lab, and eye-tracking experiments.

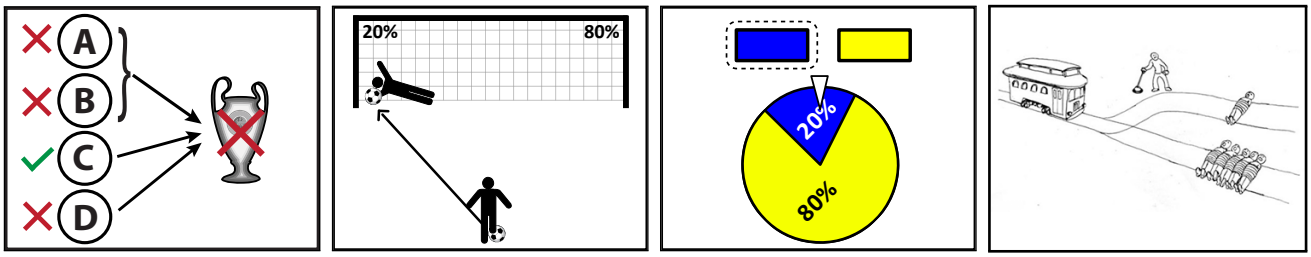
Studying responsibility judgments is challenging because they draw on such detailed knowledge about the world and the people in it. I address this challenge by bringing together two different strands of research. On the one hand, psychological theories have identified a number of factors that influence people's judgments, such as whether the action was intended or accidental. However, their lack of formal precision leaves open many possible interpretations of how exactly the postulated factors work. On the other hand, computational frameworks, such as causal Bayes nets, give us a language to describe people's causal knowledge of the world, and a calculus to determine the consequences of hypothetical interventions on the world. While formally precise, this framework lacks much of the conceptual richness that is required to adequately capture people's judgments. For example, people's ability to simulate possible worlds goes beyond the way in which this framework handles counterfactuals. We can not only imagine how particular events might have been different, we can also consider replacing a person with someone else, and then simulating how that person would have acted in the same situation.

My goal is to develop a comprehensive computational theory of responsibility that combines the best of both worlds. A theory that is formally precise *and* conceptually rich. Below, I outline what steps I have taken toward that goal. In the first section, I will show how extensions to the causal Bayes net framework that I have developed, help us better understand how people assign responsibility. To understand how people attribute responsibility, we need a better account of how they make causal judgments. In the second section, I will present a novel computational model which accurately predicts people's causal judgments. Causal judgments, in turn, are intimately linked to counterfactual simulations. I will show in the third section, how our causal knowledge of the world shapes our beliefs about what could have happened.

How we hold others responsible.

To hold someone responsible, we need to consider what causal role their action played in bringing about the outcome, and what we can infer about the person from their action (Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015). I will discuss both questions in turn.

Causal attribution: From actions to outcomes. People's responsibility judgments to individuals in groups are sensitive to the way in which individual contributions combine to determine the group outcome. Intuitively, in order to be held responsible, a person's action must have made a difference to the outcome. In the law, the but-for test is commonly employed to determine questions of factual causation (Lagnado & Gerstenberg, in press). Would the outcome have been different *but for* the person's action? However, this simple counterfactual test fails when outcomes are overdetermined, and this is often the case when several causes contribute to an outcome. Even though an individual vote is almost never pivotal for the



(a) Complex causal structure **(b)** Unexpected save **(c)** Unexpected guess **(d)** Trolley dilemma

Figure 1: Holding agents responsible in different situations. (a) Complex causal structure in which the group outcome is determined as a combination of the individual contributions. In order for the team to succeed, player C and D have to succeed as well as at least one out of A and B. (b) The goalkeeper decided to jump in the unexpected direction and saved the shot. (c) The player correctly predicted that the spinner would land on the unexpected outcome. (d) Hank decided to throw the switch in order to redirect the trolley.

outcome of an election, each voter still bears some responsibility for the outcome (Gerstenberg, Halpern, & Tenenbaum, 2015). A modified counterfactual test is required to capture this intuition. Rather than just looking at whether the person's action was pivotal in the actual situation, we need to consider whether the person's action could have been pivotal if the situation had turned out differently. The more distant the actual situation was from a situation in which the person's action would have been pivotal, the less responsibility the individual bears for the outcome.

In Gerstenberg and Lagnado (2010), we demonstrated that this modified counterfactual model captures participants' responsibility judgments. Participants' judgments were sensitive to whether individual contributions combined additively, conjunctively, or disjunctively to determine the group outcome. Individuals whose action was closer to having been pivotal were held more responsible. In Zultan, Gerstenberg, and Lagnado (2012) and Lagnado, Gerstenberg, and Zultan (2013), we looked at a wider variety of situations and found that individuals are sometimes held more responsible even though their contribution was actually *less* pivotal. Rather than just considering how close a person's action was to having been pivotal *ex post*, people also care about how critical the person's action was for a successful group outcome *ex ante*. For example, in Figure 1a, we know based on the causal structure of the task alone, that players C and D will be more critical for the group outcome than players A and B. Accordingly, participants blamed player D more than player A (when all players except C failed in their task) even though both players were equally close to having been pivotal. A model that takes into account both criticality and pivotality, accurately explains participants' responsibility judgments across a range of causal structures and performance patterns.

Dispositional inference: From actions to persons. Responsibility judgments are influenced by expectations. A critical player in a group task is expected to try as hard as possible to succeed, and a critical component in a mechanistic device is expected not to break (Lagnado & Gerstenberg, 2015). Generally, we hold others more responsible when their actions violated our expectations. However, actors sometimes receive more responsibility for expected actions. If a person chooses the action that was more likely to succeed they are praised more than when they chose an inferior option. This leaves us with a puzzle: when do unexpected versus expected actions lead to more responsibility? In Gerstenberg, Ullman, Kleiman-Weiner, Lagnado, and Tenenbaum (2014) and Gerstenberg et al. (under review), we developed a model that solves this puzzle. Rather than going directly from action expectations to judgments of responsibility, the model assumes that people first infer what the action reveals about the person's skills. Via Bayesian inference, the model updates its prior belief about what kind of person the actor is, to a posterior belief after having observed their action. This change in belief translates into an updated expectation about how the person will behave in the future. We give others credit for positive outcomes if their action improved our expectations about their future behavior. We blame others for negative outcomes, if their action lowered our expectations of them.

To see how this works, consider a goalkeeper who knows about a player's general tendency to shoot a penalty toward the right post (see Figure 1b). However, this time, unexpectedly, the player shot the ball

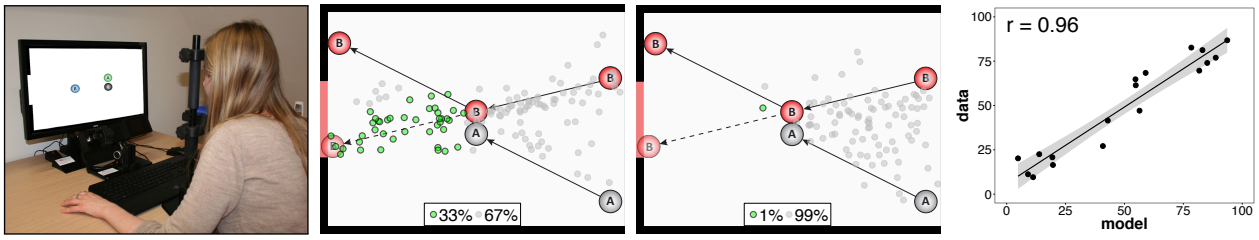
toward the left post but the goalie still saved the shot. Contrast this with a situation in which a game show contestant correctly predicted that a spinner would land on the unexpected color (see Figure 1c). Even though the goalie's and the contestant's actions were both unexpected, they lead to different inferences. While the goalie's unexpected success is consistent with him having skillfully anticipated the shot, the game show contestant was lucky but foolish. In line with our account, participants attributed more credit to goalies who saved *unexpected* shots, and more credit to game show contestants who correctly predicted that the spinner would land on the *expected* color. As further predicted by the model, actors are held more responsible when their action was pivotal (Allen, Jara-Ettinger, Gerstenberg, Kleiman-Weiner, & Tenenbaum, 2015).

When holding others responsible, we care about what they intended. Because people don't wear their intentions on their sleeves, we have to infer them from their actions. This inference is challenging since the world is noisy – sometimes good intentions lead to bad outcomes. In Gerstenberg, Lagnado, and Kareev (2010), we showed that when people assign responsibility, they care more about the intended outcome rather than what actually happened. Using an economic game that featured groups of participants interacting in pairs, we showed in Schächtele, Gerstenberg, and Lagnado (2011) that players try to deceive others about their selfish intentions in order to reduce punishment. In Kleiman-Weiner, Gerstenberg, Levine, and Tenenbaum (2015), we demonstrated how a counterfactual model that infers what outcomes an agent intended, explains people's moral permissibility judgments for a wide array of moral dilemmas (Figure 1d).

Future directions. Responsibility doesn't simply diffuse equally between the causes involved. The causal structure of the situation, our expectations about others' actions, and the intentions we infer guide our allocation of responsibility. Many large-scale problems, such as global warming, are problems of responsibility. In future work, I aim to explore experimental settings that mimic these large-scale problems – settings in which individuals occupy different roles, make their contributions sequentially, with only partial knowledge about other people's actions (cf. Gerstenberg & Lagnado, 2012). This work will help to discern what kinds of situations make people feel responsible. Further, I will look at the motivating force that anticipated attributions of blame and credit by others have for the decisions we make, and the regret we feel when things went wrong. To explore the role of norms and expectations more thoroughly, I will look at how responsibility attributions develop (Koskuba, Schlottmann, Gerstenberg, Gordon, & Lagnado, under review), and change as our understanding of others improves.

How we make causal judgments. Knowing how the world works is essential for achieving our goals (Bramley, Gerstenberg, & Lagnado, 2014; Meder, Gerstenberg, Hagmayer, & Waldmann, 2010). Philosophers have debated the nature of causality for centuries. Out of these debates, two main approaches for understanding causation have emerged. According to *process theories*, causes transfer some physical quantity (such as momentum) to effects via a spatio-temporally continuous process. According to *dependence theories*, causes are difference-makers, where the effect wouldn't have happened without the cause. Inspired by these framework theories, psychological research on how people make causal judgments has generated mixed results. Sometimes participants seem to mostly care about counterfactual dependence, whereas at other times participants are sensitive to the causal process by which the effect came about. We have developed a *counterfactual simulation model* (CSM) of causal judgment that unifies process and dependence theories of causation. It is the first model to yield quantitative predictions about how people make causal judgments for particular events.

A counterfactual simulation model of causal judgment. The CSM predicts that people's causal judgments are ultimately about difference-making. However, there are several ways in which a cause (C) can make a difference to an effect (E). For example, C can influence *whether* and *how* E occurs. Dependence theories have traditionally focused on the *whether*, while process theories have focused on the *how*. The CSM captures these aspects of causation through different counterfactual operations. To test for *whether-causation*, the model simulates what would have happened if C had been removed from the scene. The



(a) Eye-tracking **(b)** Saccades in *causal* condition **(c)** Saccades in *outcome* condition **(d)** Model fit

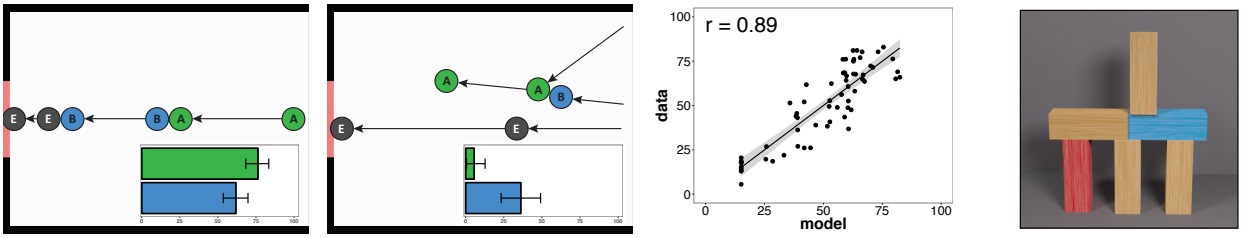
Figure 2: (a) Tracking participants' eye-movements as they view dynamic physical animations. Figures (b) and (c) show what paths ball A and B traveled on before the collision, B's actual path after the collision, and the path that B would have taken (dashed line) if ball A had been removed from the scene. On top of that, the figures show the endpoints of participants' saccades in the interval between when the balls entered the scene and when they collided. Green dots indicate counterfactual looks along B's counterfactual path, gray dots indicate other looks. Participants in the *causal* condition spontaneously simulated where ball B would have gone much more so than participants in the *outcome* condition. Figure (d) shows that the counterfactual simulation model captures participants' causal judgments to a high degree of quantitative accuracy.

more certain the model is that the outcome would have been different, the more of a whether-cause C was. To test for *how-causation*, the model simulates what would have happened if C had been somewhat different, rather than removing it from the scene. The model further considers whether C was *sufficient* and *robust* in bringing about E.

The CSM can be applied to any domain which can be expressed with a generative model that affords counterfactual manipulations. So far, we have tested the CSM in a series of experiments in which we presented participants with physically realistic animations of colliding billiard balls (see Figure 2). Depending on the outcome of the clip, participants indicated whether they believed that ball A caused ball B to go through the gate, or prevented it from going through (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012, 2014). Another group of participants was asked to indicate whether ball B would have gone through the gate if ball A hadn't been present in the scene. The results revealed a very close correspondence between participants' counterfactual and causal judgments. By linking causal judgments to graded beliefs about counterfactual outcomes, the CSM is the first model to provide a quantitatively adequate account of people's causal judgments about particular events. The more certain participants were that ball B would have missed the gate in the absence of ball A, the more they said that A caused B to go through the gate (Figure 2d).

Eye-tracking causality. The close quantitative correspondence between counterfactual and causal judgments provides empirical support for the CSM. The CSM also makes a very strong process-level prediction: people use mental simulation to arrive at their causal judgment. To test this prediction, we ran a study in which we tracked participants' eye-movements (Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, resubmitted; see Figure 2a). Between three conditions, we manipulated what question participants were asked. In addition to the causal and counterfactual conditions described above, we also included a condition in which participants were asked to make a judgment about the actual outcome (e.g. whether B completely missed the gate). Figure 2b and c show the result of classifying participants' saccades in the causal and outcome condition, respectively. In the causal condition, participants were much more likely to spontaneously simulate where ball B would go if ball A hadn't been present in the scene. These results were confirmed by a hidden Markov model which yields a dynamic landscape of how participants divide their attention over time. Even though counterfactual processing has been postulated for many aspects of cognition, to our knowledge this is the first study to directly show spontaneous counterfactual simulation in the service of causal judgment.

Multiple causes. To investigate more closely to what extent participants' causal judgments are sensitive to the different aspects of causation that the CSM postulates, we created clips that featured interactions between two candidate causes A and B, and a target ball E. For example, in a causal chain (Figure 3a), participants say that ball A was somewhat more causally responsible for E's going through the gate than



(a) Causal chain **(b)** Double prevention **(c)** Model fit **(d)** A tower of blocks

Figure 3: (a) and (b) show diagrams of clips that are modeled after situations often discussed in the philosophical and psychological literature on causation. For example, diagram 2 depicts *double prevention*, where ball B prevents ball A from preventing ball E from going through the gate. The bar charts in each diagram show participants' mean ratings with confidence intervals. (c) The scatter plot illustrates the relationship between model prediction and mean judgments for the full experiment. (d) The red and blue blocks carry some responsibility for the tower's stability.

ball B. However, note that B's rating is much greater than zero even though ball E would have still gone through the gate even if ball B had been removed. As the CSM predicts, participants not only care about *whether* the outcome came about, but also *how* it did. If ball B's initial position had been somewhat perturbed, then E would have gone through the gate differently from how it actually did. Accordingly, both balls A and B were how-causes of E's going through the gate. Because A was also a whether-cause, it is seen as more causal than B. Considering how-causation also helps explain why ball B's rating is relatively low in the case of double prevention (Figure 3b). Even though E wouldn't have gone through the gate if B hadn't knocked ball A out of the way, B didn't directly affect *how* E went through the gate.

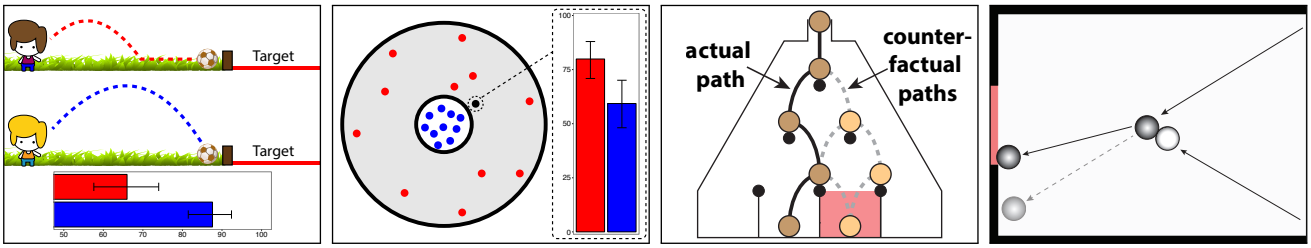
With only three billiard balls, we can recreate the majority of thought experiments that have been discussed in the philosophical literature on actual causation, including causal chains, double prevention, overdetermination, preemption, joint causation, etc. The CSM explains people's judgments across all of these situations (Figure 3c), as well as situations that involve barriers preventing balls from going down certain paths, or teleports that beam balls from one location of the screen to another.

Future directions. We use a wide variety of causal terms to describe what happened. However, most research to date has only looked at "caused" and "prevented". I will use the different aspects of causation that the CSM highlights as a tool for exploring the semantics of different causal terms such as "caused", "enabled", or "helped". By incorporating normative expectations, the CSM will be able to deal with cases of causation by omission, which are related to the semantics of "let".

Eye-tracking is an excellent tool for studying mental simulation. I am excited about using the methods I have developed to study how counterfactual simulation develops in children. As an implicit measure, eye-tracking circumvents the difficulties children have with verbally processing counterfactual questions.

So far, I have used the CSM to explain causal judgments about dynamic collision events. I will explore the CSM in novel domains that span across and go beyond physical reasoning. For example, to determine the extent to which the blue and red blocks carry responsibility for the tower's stability in Figure 3d, it seems natural to mentally simulate what would happen if each block was removed from the scene. By creating clips similar to the ones used by Heider and Simmel, I will explore people's judgments about intentional agents.

How we bring order into counterfactual worlds. The philosopher David Lewis tried to reduce causality to statements about counterfactuals whose truth was grounded in a possible world semantics. In this account, the notion of similarity between possible worlds was used to determine which counterfactual statements are true or false. Since Lewis aimed to provide a reductive account of causation, he had to define the similarity between different possible worlds in non-causal terms. This proved to be a difficult (if not impossible) task. The question of how we bring order into counterfactual words is still largely unresolved. In my work, I have shown how people draw on their causal knowledge to answer counterfactual questions (Gerstenberg, Bechlivanidis, & Lagnado, 2013). By characterizing people's knowledge in terms of rich generative models (such as their intuitive understanding of physics), we can test more precisely what makes some possible worlds lie closer than others (Gerstenberg & Goodman, 2012; Gerstenberg &



(a) Almost reached the target **(b)** Almost hit the center **(c)** Almost won the game **(d)** Almost caused the ball to go through the gate

Figure 4: Different situations in which counterfactual worlds came close. (a) The boy almost threw his ball into the target region. (b) The black dart almost hit the center. (c) The marble almost landed in the red winning region. (d) The white ball almost caused the black ball to go through the gate.

Tenenbaum, in press; Goodman, Tenenbaum, & Gerstenberg, 2015).

We have shown how people's judgments about whether something "almost happened" provides a window into their counterfactual representations (Gerstenberg & Tenenbaum, 2016). People judge that something almost happened when a small perturbation to the cause would have been sufficient to bring about the counterfactual outcome. For example, even though the balls of the top and bottom character in Figure 4a landed equally close to the target, only the bottom character's ball is judged to have almost reached the target region. Prior expectations also affect what worlds we consider close. We are more willing to say that the black shot in Figure 4b almost hit the center when it came from the red player (who had been doing poorly thus far) than from the blue player. We also looked at whether people's judgments of "almost" are influenced by the different ways in which the counterfactual outcome could have come about, and found that only the distance to the closest possible world matters (Figure 4c). Drawing on the CSM, we developed an account of what it means for one event to have "almost caused" another (Figure 4d). Finally, we have shown how norms influence the availability of counterfactuals which consequently affect people's causal judgments (Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015).

Future directions. My explorations into counterfactual closeness have only just begun. In the future, I will look more closely at the way in which changes in expectations over time (such as momentum in sports) influence which counterfactual worlds move closer. I will also explore what role people's motivations play in how they construe the distance to possible worlds. An observer's motives may bias their beliefs about what would have happened in the absence of the cause, as well as their perception of how close the actual outcome was (cf. Figure 4d). Generally, we'd rather "almost win" than "lose".

Future work. Moving forward, I will continue to explore the question of how we hold others responsible, as well as the processes of causal judgment and counterfactual simulation that form the basis of our responsibility judgments. In my work so far, I have mostly looked into situations in which people have relatively little uncertainty about how the world works. Our everyday life, however, is replete with uncertainty. In future work, I aim to look more at how people's understanding of a situation develops over time, and how this improved understanding affects their judgments (cf. Bramley, Gerstenberg, & Tenenbaum, 2016). I am particularly interested in investigating the relationship between social learning and responsibility judgments. How do we come to understand what others are like, and how does our knowledge of their strengths and weaknesses affect our evaluations of their behavior? I am excited about expanding the computational modeling techniques I have used so far in my research, and to continue developing novel experimental paradigms that allow me to rigorously test these ideas.