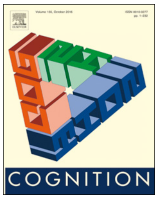
Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit



Original Articles

Lucky or clever? From expectations to responsibility judgments

Tobias Gerstenberg[a,*], Tomer D. Ullman[a], Jonas Nagel[b], Max Kleiman-Weiner[a], David A. Lagnado[c], Joshua B. Tenenbaum[a]

[a] *Massachusetts Institute of Technology, United States*
[b] *Göttingen University, Germany*
[c] *University College London, United Kingdom*

ARTICLE INFO

Keywords:
Responsibility
Causality
Expectations
Counterfactuals
Pivotality

ABSTRACT

How do people hold others responsible for the consequences of their actions? We propose a computational model that attributes responsibility as a function of what the observed action reveals about the person, and the causal role that the person's action played in bringing about the outcome. The model first infers what type of person someone is from having observed their action. It then compares a prior expectation of how a person would behave with a posterior expectation after having observed the person's action. The model predicts that a person is blamed for negative outcomes to the extent that the posterior expectation is lower than the prior, and credited for positive outcomes if the posterior is greater than the prior. We model the causal role of a person's action by using a counterfactual model that considers how close the action was to having been pivotal for the outcome. The model captures participants' responsibility judgments to a high degree of quantitative accuracy across three experiments that cover a range of different situations. It also solves an existing puzzle in the literature on the relationship between action expectations and responsibility judgments. Whether an unexpected action yields more or less credit depends on whether the action was diagnostic for good or bad future performance.

1. Introduction

In the quarter final of the 2006 FIFA World Cup, the Germany versus Argentina match came down to penalty shots. Unbeknownst to the Argentinian team, the German goalkeeper, Jens Lehmann, was handed a piece of paper that indicated where each of the Argentinian players was likely to shoot. Lehmann ended up saving two penalties, and the German team won the game. Clearly, Lehmann deserves credit for the team's win. But how much, and on what grounds?

Let us suppose that the following took place: Lehman was told that the first shooter often aims the ball at the left corner. Lehmann jumped to this corner and saved the ball. For the second shooter, Lehmann was told again to expect a shot in the left corner. However, this time Lehmann jumped in the opposite corner, and again saved the shot, even though his opponent kicked the ball in the unexpected direction. Would you give Lehmann more credit for the first, or the second save? And suppose Lehmann had failed to save both shots. Would you have blamed him more for failing to save the shot that went in the expected direction, or the unexpected one?

In this paper, we investigate how people hold others responsible for their actions. Most existing accounts predict that unexpected actions elicit greater attributions of responsibility than expected actions (Brewer, 1977; Fincham & Jaspars, 1983; Malle, Guglielmo, & Monroe, 2014; Petrocelli, Percy, Sherman, & Tormala, 2011), and, more generally, that unexpected events are more likely to be cited as the cause of an outcome (Halpern & Hitchcock, 2015; Hart & Honoré, 1959/1985; Hilton & Slugoski, 1986; Hitchcock & Knobe, 2009; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015). However, recently Johnson and Rips (2015) reported a series of experiments in which participants held agents *more* responsible when positive outcomes resulted from *expected* actions. In their experiments, an agent faced a choice between multiple options that differed in their probability of bringing about a positive outcome. They found that participants held the agent more responsible for a positive outcome when the agent chose an option that was better than any of the alternatives, and less responsible when the agent chose an inferior option.

Together, these findings present a puzzle: When do we assign more responsibility for unexpected actions (as most theories predict), and when do we assign less responsibility? We present a computational model that solves this puzzle. The model relies on two processes: the first process is a *dispositional inference* that captures what an action reveals about a person. Specifically, we propose that a person will be credited (or blamed) to the degree that their action reveals they are the sort of person who will get things right (or wrong) in the future. To go

[*] Corresponding author at: MIT Building 46-4053, 77 Massachusetts Avenue, Cambridge, MA 02139, United States.
E-mail address: tger@mit.edu (T. Gerstenberg).

https://doi.org/10.1016/j.cognition.2018.03.019
Received 4 December 2017; Received in revised form 21 March 2018; Accepted 23 March 2018

back to our opening example, Lehmann will be credited more for saving the unexpected shot because we infer by that action that Lehmann is a skilled goalie. However, if Lehmann chose an unexpected action in a pure game of chance, this would be diagnostic of poor future performance and so our model predicts little credit in this case.

The second process is a *causal attribution* of the role that a person's action played in bringing about the outcome. People are held more responsible to the extent that their action was pivotal in bringing about the outcome.

Our formal framework for explaining responsibility judgments draws on a rich literature in attribution theory, as well as recent work on modeling causal judgments. We briefly review each of these strands of research, focusing on the aspects that are most relevant for our framework. We then present our computational model in detail, and subsequently test the fine-grained predictions of our model in three experiments that vary action expectations, and the extent to which a person's action made a difference to the outcome. We discuss how our model relates to previous work, and how different comparison standards may affect judgments of responsibility. We conclude by highlighting future avenues of research motivated by the model and results presented here.

1.1. Dispositional inference: from actions to persons

Early attribution theorists proposed Bayesian inference as a normative framework for making diagnostic inferences about a person from observing their actions (Ajzen, 1971; Ajzen & Fishbein, 1975, 1978; Fischhoff & Beyth-Marom, 1983; Fischhoff & Lichtenstein, 1978; Morris & Larrick, 1995; Trope, 1974; Trope & Burnstein, 1975). For the Bayesian framework to support inferences from observed variables (behavior) to latent variables (mental states), it requires a model that captures how the latent and observed variables relate. Essentially, in order to assign responsibility to others, we need a model of decision-making that expresses how we believe people make choices based on their mental states. A key assumption for making sense of other people's behavior in this way is the *principle of rational action* (Dennett, 1987). It states that a person chooses an action that is expected to achieve a desired goal in the most efficient way, subject to the person's beliefs and abilities (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009; Gilbert, 1998; Goodman et al., 2006; Heider, 1958; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Malle & Knobe, 1997; Pantelis et al., 2014; Wellman & Bartsch, 1988).

To the extent that a person acts in line with our expectations, we do not learn much beyond what we already know, and need not update our beliefs. However, when a person's action violates our expectation then we need to make sense of their behavior, either finding situational factors that influenced their actions, or updating our beliefs about who they really are (Duff, 1993; Frieze & Weiner, 1971; Koster-Hale & Saxe, 2013; Uhlmann, Pizarro, & Diermeier, 2015; Weiner, 1985; Weiner, Heckhausen, Meyer, & Cook, 1972). Did the agent have some special skill and behave optimally in light of having this ability, or did the agent lack the relevant skill, and the positive outcome was the lucky result of poor decision-making (cf. Morse, 2003; Rachlinski, 2002–2003; Sinnott-Armstrong & Levy, 2011; van Inwagen, 1978)?

Our model predicts that attributions of responsibility are closely linked to our expectations. We credit a person if their action indicates that they are better than a comparison standard. Conversely, we blame a person if their action reveals that they are worse than we expected.

1.2. Causal attribution: from actions to outcomes

Research on causal attribution has identified a host of factors that influence people's causal judgments (Einhorn & Hogarth, 1986; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012, 2014, 2015; Lagnado, Waldmann, Hagmayer, & Sloman, 2007; Sloman, 2005; Sloman & Lagnado, 2015; White, 2014; Wolff, 2007). In order to be held responsible for an outcome, a person's action must be causally connected to the outcome. We predict that the extent to which a person is blamed or credited for an outcome depends on the perceived causal influence that their action had on the outcome. To determine what role an action played in bringing about the outcome, we need a causal model of the situation that captures how the action of interest and other candidate causes affected the outcome. Here, we take inspiration from work in philosophy (Woodward, 2003; Yablo, 2002) and computer science (Halpern & Pearl, 2005; Pearl, 2000) that models causal relationships in terms of counterfactual contrasts over a causal model of the situation.

Within this framework, a variable qualifies as a cause of an outcome if the outcome would have been different had the variable taken on a different value (Lewis, 1973). However, this test of counterfactual dependence runs into problems when outcomes are overdetermined by multiple, individually sufficient causes. For example, in elections, the outcome would often not have been any different if a single voter had changed her mind. However, we still want to say that each voter has some degree of responsibility for the outcome. Halpern and Pearl (2005) proposed a structural model of causal attribution that handles this and other problems by replacing the simple counterfactual test of causation with a test of counterfactual dependence under contingency. A variable can qualify as a cause even when it did not make a difference in the actual situation, as long as there was a possible situation that could have arisen, in which the event would have made a difference.[1] Chockler and Halpern (2004) have proposed that the closer a person's action was to having been pivotal, the greater their causal responsibility for the outcome. Prior research has shown that pivotality is an important factor in how people attribute responsibility (Gerstenberg, Halpern, & Tenenbaum, 2015; Gerstenberg & Lagnado, 2010, 2012, 2014; Lagnado & Gerstenberg, 2015; Lagnado, Gerstenberg, & Zultan, 2013; Wells & Gavanski, 1989; Zultan, Gerstenberg, & Lagnado, 2012).

In this paper, we will look at relatively simple settings in which a decision-maker chooses between two options. In some of the situations, their actions turn out to be pivotal – the outcome would have been different if they had acted differently – whereas in other situations, their actions aren't pivotal – the outcome would have been the same even if they had chosen the other option. We predict that a person is viewed as more responsible for an outcome when her action was pivotal.

2. Computational model

Our model assigns a degree of responsibility to people making decisions under uncertainty that result in positive or negative outcomes (cf. Botti & McGill, 2006; Leonhardt, Keller, & Pechmann, 2011; Nordbye & Teigen, 2014; Parkinson & Byrne, 2017). Our model has two components: (i) a dispositional inference from the person's action to their character, which affects the model's expectation about the person's future behavior, and (ii) a causal inference about the relationship between the person's action and the outcome. We will discuss each component in turn.

2.1. Dispositional inference and expectation change

The first component of our model formalizes how we update our expectations about a person's future performance after having observed the person's action, and the outcome that resulted. This inference involves two steps. The first step is to update our belief about the type of person the decision maker is. The second step is to transform this new belief into an updated expectation about how well the person will do in the future. We will discuss each step in turn.

[1] Much of the work goes into specifying which contingencies are allowed when checking for whether a counterfactual dependence holds between the candidate cause and effect (cf. Livengood, 2011). In this paper, we will focus on settings in which these difficulties do not arise.

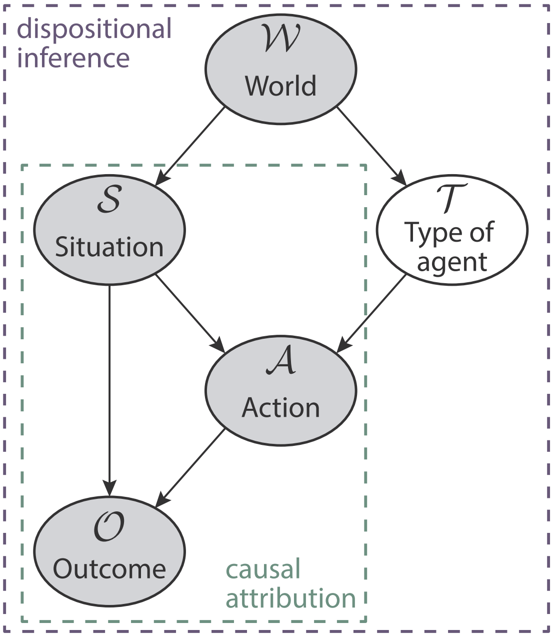


Fig. 1. Generative model of an observer who considers how much responsibility a person should receive for an outcome that resulted from the person's taking an action in a particular situation. The world determines the distribution over agent types and possible situations. Agent types differ in how they choose actions in a given situation. The outcome is a deterministic function of what action the agent took in a particular situation. The model predicts that an observer engages in two processes: (1) a *dispositional inference* over the agent type (white node) from having observed the world, situation, action, and outcome (shaded nodes), and (2) a *causal attribution* that determines the extent to which the person's action was causally responsible for the outcome.

2.1.1. Dispositional inference

We want to infer the type of a person $t \in \mathscr{T}$, from their observed action $a \in \mathscr{A}$:

$$P(\mathscr{T}=t|\mathscr{A}=a) \propto P(\mathscr{A}=a|\mathscr{T}=t)\cdot P(\mathscr{T}=t). \tag{1}$$

The prior $P(\mathscr{T}=t)$ expresses the model's initial belief about the person's type, before observing their action. The likelihood function $P(\mathscr{A}=a|\mathscr{T}=t)$ expresses how a person decides to take an action, given their type. These two pieces determine the posterior belief about the person $P(\mathscr{T}=t|\mathscr{A}=a)$. In principle, the space of agent types $\mathscr{T}$ can be made very rich by incorporating the many factors that are known to influence people's actions. Here, we focus on a small space of agents that is sufficiently rich to capture aspects of a person's personality that may be relevant for assigning responsibility in scenarios that involve achievement and failure.

Fig. 1 illustrates the inference problem that participants face in our experiments. The world $\mathscr{W}$ refers to the scenario where the action takes place. Each condition in our experiments is a different world in this sense. The world determines what kinds of situations $\mathscr{S}$ are possible, and what types of agents $\mathscr{T}$ are likely to be present in the world. Agents take actions $\mathscr{A}$ depending on the situation they find themselves in, and depending on what type of agent they are. In our experiments, we focus on simple situations that are characterized by two action alternatives. What action an agent takes in a particular situation determines whether the outcome $\mathscr{O}$ is positive or negative.

Let us illustrate how this works more concretely, via the examples shown in Fig. 2. Fig. 2a shows the "goalie world". The possible outcomes for the goalie are saving the ball ($\mathscr{O}$ = positive), or failing to save the ball ($\mathscr{O}$ = negative). Outcomes are mapped to a binary reward r, with $r = 1$ for a positive outcome, and $r = 0$ for a negative outcome. In this particular situation, the goalie knows that the striker has a 20% chance of shooting the ball towards the left corner (from the perspective of the striker), and an 80% chance of shooting the ball towards the right corner. The striker does not know that the goalie knows about his tendency to shoot towards the right. The goalie chose the action of jumping towards the unlikely 20% direction and saved the ball, as the striker decided to kick left.

Fig. 2b shows the "spinner world". Here, the possible outcomes are correctly predicting what color the spinner will land on ($\mathscr{O}$ = positive), or making a wrong prediction ($\mathscr{O}$ = negative). In this situation, the spinner had a 20% chance of landing on blue, and an 80% chance of landing on yellow. The player correctly predicted that the spinner will land on blue (the unlikely outcome).

In both examples, the observed variables are the world $\mathscr{W}$, the situation $\mathscr{S}$, the action the agent took $\mathscr{A}$, and the outcome that resulted $\mathscr{O}$. The specific agent type $\mathscr{T}$ is unobserved but can be inferred through Bayes' theorem:

$$P(\mathscr{T}=t|\mathscr{W}=w, \mathscr{S}=s, \mathscr{A}=a) = \frac{P(\mathscr{A}=a|\mathscr{S}=s, \mathscr{T}=t)\cdot P(\mathscr{T}=t|\mathscr{W}=w)}{\sum_{k\in\mathscr{T}} P(\mathscr{A}=a|\mathscr{S}=s, \mathscr{T}=k)\cdot P(\mathscr{T}=k|\mathscr{W}=w)}, \tag{2}$$

where t ranges over the different agent types $\mathscr{T}$. This equation shows that in order to infer the agent type, we need to know the distribution over agent types in a given world $P(\mathscr{T}|\mathscr{W})$, and how different agent types choose their actions in different situations $P(\mathscr{A}|\mathscr{T}, \mathscr{S})$. Let us focus on the latter quantity first.

Agent decision functions. We model a particular situation s as containing two types of signals. First, a *probability signal* s_p that represents the probability that an action will result in a positive outcome (e.g. there is 20% chance of a positive outcome if the player predicts that the spinner will land on blue, and a 80% chance if the player predicts yellow). The second signal is an *outcome signal* s_o that represents a cue about the actual outcome of a specific case (e.g. the force with which the spinner is spun reveals that it will land on blue). We define three agent types that differ in the signals they have access to, and how they make use of the information. The ability to detect and correctly use the outcome signal is a way of formalizing an agent's skill.

The *average agent* makes its choice based on the probability signal. The *skilled agent* has access to the outcome signal, and correctly anticipates what will happen. Finally, the *unskilled agent* also has access to the outcome signal. However, it uses the outcome signal in the wrong way and is more likely to choose the action that yields no reward.

More formally, all agents choose their actions according to a softmax decision rule. The softmax decision function is a standard choice rule for modeling an agent's planning and decision-making in uncertain environments (cf. Luce, 1959; Sutton & Barto, 1998). The probability that an *average agent* chooses a specific action a is

$$P(a|s_p, \mathscr{T}=\text{average}) = \frac{exp(\beta\cdot\hat{r}_a)}{\sum_{k\in\mathscr{A}} exp(\beta\cdot\hat{r}_k)}, \tag{3}$$

where $\hat{r}_a$ is the expected reward that results from taking action a, and β is a noise parameter. To normalize, the denominator sums over all possible actions. The expected reward of each action is determined by the probability signal s_p. For example, in the situation shown in Fig. 2a, the expected reward for jumping in the left corner (again, from the striker's perspective) is $\hat{r}_{\text{left}} = r\cdot s_p = 1\cdot 0.2 = 0.2$, and for jumping right it is $\hat{r}_{\text{right}} = 1\cdot 0.8 = 0.8$.[2]

The β parameter interpolates between random choice and expected reward maximization. If β is large, then the agent almost always chooses the action with the greatest expected reward. If $\beta = 0$, the agent chooses an action at random. For intermediate values of β, the agent chooses actions in proportion to their expected reward. The

[2] Note that in the model, we assume that agents treat the probabilities in the spinner and goalie setup identically. In the real world, we might place more uncertainty on the probabilities in the goalie case (where a striker can easily override their past tendency) compared to the spinner case.

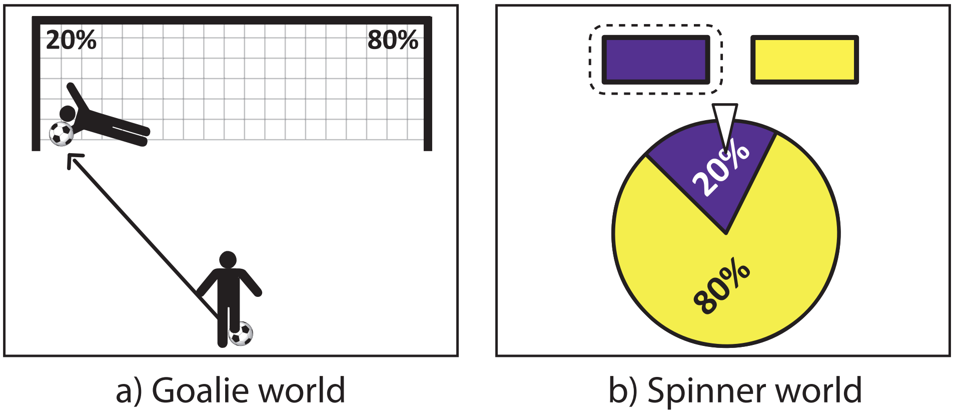


Fig. 2. Examples of stimuli shown to participants in the (a) goalie condition, and (b) spinner condition of Experiments 1 and 2. The goalkeeper in (a) and the game show contestant in (b) chose the unexpected action and the outcome was positive.

decision noise parameter β captures any uncertainties in the environment that we do not model explicitly, but that could potentially affect the agent's decision-making.

The *skilled agent* is able to anticipate what will happen in a given situation. For example, irrespective of what a penalty-taker's general tendency is, the skilled goalie will be able to anticipate where the penalty-taker will shoot the ball this time. The skilled agent has access to the outcome signal s_o which reveals what the true reward of each action will be, r_a^{true}. However, just like the average agent, the skilled agent's decisions are noisy. One way to think about the decision-noise here is that it captures the agent's uncertainty about the outcome signal it received. The probability that the skilled agent chooses a certain action *a* is

$$P(a|s_o, \mathscr{T} = \text{skilled}) = \frac{exp(\beta \cdot r_a^{\text{true}})}{\sum_{k \in \mathscr{A}} exp(\beta \cdot r_k^{\text{true}})}, \tag{4}$$

where r_a^{true} is the actual reward that will result from taking action *a*.[3]

The *unskilled agent* also has access to the outcome signal s_o. However, the unskilled agent gets things the wrong way around and uses the signal incorrectly. Rather than choosing randomly, an unskilled agent is actually more likely to choose the action that will result in lower reward. Because it is worse than random, we will sometimes refer to the 'unskilled agent' as the 'bad agent'. The probability that the unskilled agent chooses a certain action *a* is

$$P(a|s_o, \mathscr{T} = \text{unskilled}) = \frac{exp(\beta \cdot (-r_a^{\text{true}}))}{\sum_{k \in \mathscr{A}} exp(\beta \cdot (-r_k^{\text{true}}))}. \tag{5}$$

In our experiments, we informed participants about the different agent types and how they make their decisions. This space of agents is rich enough to capture important aspects, such as skill and rational decision-making, that are relevant for how people evaluate another person's actions in an achievement context. At the same time, the space of agents is small enough to be assessed in full by explicitly asking participants what agent they believe a person was, after having observed the person's action.

Prior over agent types. There are many factors that can influence what sort of agents people consider plausible in a given world. For example, people may think it a-priori more likely that some people possess the skill to correctly anticipate the outcome in a game like soccer, compared to a game like the spinner prediction task. In Experiments 2 and 3, we introduce participants to the three different agent types. Instead of explicitly probing participants' beliefs about how likely they consider each agent type to be a-priori, we infer participants' priors based on their judgments of what type of agent they think a person is, after having observed the person's action.

2.1.2. Expectation change

So far we have shown how the model updates a prior belief over the agent type, conditioned on the agent's action in a particular situation. But how does this dispositional inference influence responsibility judgments? We propose that dispositional inferences are relevant to the extent that they change our expectations about how the agent will behave compared to our prior expectations. The second step in our model captures this intuition.

Before observing the particular agent's action, the prior expectation of how much future reward an agent will receive in a given world is

$$\mathbb{E}[r|\mathscr{W} = w] = \sum_{s \in \mathscr{S}} \sum_{t \in \mathscr{T}} \sum_{a \in \mathscr{A}} r_a \cdot P(a|s, t) \cdot P(t|w) \cdot P(s|w), \tag{6}$$

where we sum over the different situations $\mathscr{S}$ that may arise, the different agent types $\mathscr{T}$, and the different actions $\mathscr{A}$ each agent might take. This prior expectation about how an agent will do is then contrasted with the posterior expectation of reward after observing a particular agent take action a_{obs} in situation s_{obs}

$$\mathbb{E}[r|\mathscr{W} = w, s_{\text{obs}}, a_{\text{obs}}] = \sum_{s \in \mathscr{S}} \sum_{t \in \mathscr{T}} \sum_{a \in \mathscr{A}} r_a \cdot P(a|s, t) \cdot P(t|w, s_{\text{obs}}, a_{\text{obs}}) \cdot P(s|w), \tag{7}$$

where we have replaced the prior over agent types $P(t|w)$ with the posterior $P(t|w, s_{\text{obs}}, a_{\text{obs}})$. The difference between these expectations expresses how our belief about how this particular agent will do in the future has changed from our prior expectation about how well agents would do in this world:

$$\text{Difference in expected reward} = \mathbb{E}[r|\mathscr{W} = w, s_{\text{obs}}, a_{\text{obs}}] - \mathbb{E}[r|\mathscr{W} = w] \tag{8}$$

Our model predicts that an observer credits an actor for a positive outcome to the extent that the observer's expectations about the actor's future reward are greater than the prior expectations. Observers are predicted to blame actors for negative outcomes to the extent that the observer's expectations about the actor's future are lower than the prior expectations. In short, exceeding expectations for positive outcomes means more credit, whereas not meeting expectations for negative outcomes means more blame.

The predictions of the model are sensitive to where people's prior expectations come from. Specifically, it matters whether prior

[3] We note that in many situations, optimal decisions require taking into account both information about the past as well as any additional signals that are specific to the actual situation. Here, we make the simplifying assumption that the outcome signal s_o is perfectly diagnostic for what will happen, and so the optimal thing to do is to completely disregard the prior expectation. Similarly, in reality, poor decisions are not simply the result of a consistent misuse of valid cues. Our characterization is a simplification meant to broadly capture poor decision-making.

expectations are agent-specific or population-specific (cf. Sytsma, Livengood, & Rose, 2012). For example, when we consider how much Tom should be blamed for a negative outcome, we can either use a prior expectation that is specific to Tom ("How do I expect Tom to act?"), or use a more generic expectation ("How do I expect a reasonable person to act?"). How much we blame Tom depends on our comparison standard. In the experiments reported in this paper, participants only observe an agent take a single action. Hence, these experiments do not tease apart how different prior comparison standards may influence people's responsibility attributions. In the General Discussion, we will talk more about the different ways in which prior expectations may influence responsibility judgments.

2.2. Causal attribution

The *difference in expected reward* is the first component in our responsibility model. The second component captures the causal role that the person's action played for bringing about the outcome in a particular situation (see *causal attribution* in Fig. 1). Above, we have shown that one way of capturing how much a person's action affected the outcome, is by considering how close the action was to having made a difference to the outcome.

Consider a situation in which the outcome of an election was 5 to 2 in favor of candidate C_1 over candidate C_2. According to a simple counterfactual but-for test, none of the 5 C_1 votes qualifies as a cause of C_1's victory. Candidate C_1 would still have won even if any of the voters had voted for candidate C_2 instead. However, we still want to say that each of the voters was a cause of the outcome, and bears some responsibility for C_1's win. Halpern and Pearl's (2005) structural-model of causation relaxes the simple counterfactual test, by considering counterfactual dependence under different possible situations that could have arisen. For example, if the outcome of the vote above were 4 to 3 (instead of 5 to 2), then candidate C_1's win would have been dependent on each of the 4 voters. If, in this situation, any of her supporters were to vote for candidate C_2 instead, then the outcome would have been different.

Based on this modified counterfactual definition of causation, Chockler and Halpern (2004) proposed a structural model of responsibility that links the degree to which a cause is held responsible for an outcome to the minimal distance between the actual situation and the counterfactual situation in which the candidate cause would have been pivotal. Accordingly, the responsibility (or *pivotality*)[4] of a person taking action $a \in \mathscr{A}$, for a particular outcome $o \in \mathscr{O}$ in a situation $s \in \mathscr{S}$ is defined as

$$\text{Pivotality for outcome}(\mathscr{A} = a, \mathscr{O} = o, \mathscr{S} = s) = \frac{1}{N+1}, \tag{9}$$

where N is the minimal number of changes that is needed to establish a counterfactual dependence between the person's action a and the outcome o in situation s. Applying this model to the voting example above yields a responsibility of $\frac{1}{2}$ for each of the five voters who voted in favor of candidate C_1. While none of the voters made a difference to the outcome in the actual situation, each of them would have made a difference in a situation in which one of the other four voters had voted for candidate C_2 instead. A single change ($N = 1$) is required to make any supporting voter pivotal for C_1's win. In previous work, we have shown that people's judgments of responsibility are sensitive to how close a person's contribution was to being pivotal for the outcome (Allen, Jara-Ettinger, Gerstenberg, Kleiman-Weiner, & Tenenbaum, 2015; Gerstenberg et al., 2015; Gerstenberg & Lagnado, 2010, 2014; Lagnado & Gerstenberg, 2015, 2017; Lagnado et al., 2013; Zultan et al., 2012).

In some of the worlds we consider below (the goalie and the spinner worlds) the agent's action is always pivotal. If a goalkeeper did not save the ball, that means she would have saved the ball if she had jumped in the other direction. However, in other worlds, we consider situations in which the outcome would have been positive (or negative) no matter what the agent did. For example, in one experiment a gardener faces the choice of which fertilizer to use in order to make a flower grow. In some situations, in turns out that the flower would have grown no matter which of the fertilizers the gardener chose. In these situations, the agent's pivotality is reduced. We predict that a person will be judged more responsible for an outcome when her action was pivotal in the actual situation. In line with this prediction, work on people's judgments about dynamic physical interactions has shown that the extent to which a candidate cause is perceived as having been pivotal strongly affects causal judgments (Gerstenberg et al., 2012; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2014, 2015; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Gerstenberg & Tenenbaum, 2016, 2017; Stephan, Willemsen, & Gerstenberg, 2017).

2.3. Putting it together: responsibility judgments

We predict that judgments of responsibility are sensitive to what the observer learned about the person from their action ('Difference in expected reward'), and the extent to which the person's action made a difference to the outcome ('Pivotality for outcome'). For simplicity, we assume that both factors of the model combine additively to affect judgments of responsibility.[5]

$$\begin{aligned}\text{Responsibility} = \alpha &+ w_1\cdot(\text{Difference in expected reward})\\ &+ w_2\cdot(\text{Pivotality for outcome}),\end{aligned} \tag{10}$$

where w_1 and w_2 determine how strongly each component of the model influences participants' responsibility judgments, and α is a parameter that allows us to map from the model's scale to participants' response scale.

Our model has four free parameters in total. β captures the decision noise in the action selection functions of the different agents (cf. Eqs. (3)–(5)). We fit $\beta = 1.5$ by minimizing the sum of squared differences between the model's predicted posterior over the different agent types, and participants' empirical judgments. We allowed participants' priors over the three different agents to vary between experiments. These priors were inferred from participants' posterior judgments and constrained by the decision noise parameter β. The remaining three parameters ($\alpha = 6$, $w_1 = 28$, and $w_2 = 28$) capture how 'Difference in expected reward' and the 'Pivotality for outcome' are weighted and mapped onto participants' response scale (cf. Eq. (10)).[6] We used the same set of parameters to derive predictions for each individual experiment. Finally, we used cross-validation methods to compare our model with lesioned models that consider only one of the two aspects, as well as an alternative model that makes predictions based on action expectations directly.

In the remainder of this paper, we report the results of several experiments that test this model of responsibility attribution for actions under uncertainty. Experiment 1 shows that the mapping between action expectations and responsibility judgments differs depending on the context of the scenario. Whether unexpected positive outcomes result in more or less credit depends on the plausibility of agents exhibiting skill.

[4] We use the term *pivotality* here instead of *responsibility* because in our model, pivotality features as one of the factors that is predicted to influence how people assign responsibility.

[5] While we assume an additive relationship here, we believe that in order for a person to be held responsible for the outcome, their action must have made a difference to the outcome in one way or another. If there was no possible situation in which a person's action could have been pivotal, then we would not expect the person to be held responsible for the outcome. Such boundary cases do not arise for the experiments we discuss below.

[6] We normalized each predictor because their ranges differ. The fact that both predictors have the same weight suggests that, on average, participants in our experiments cared about both factors to similar extents.

Experiment 2 replicates the results of Experiment 1 and tests the predictions of our model by explicitly informing participants about the different agent types. Participants blame agents for actions that are indicative of poor decision-making, and credit agents for actions that suggest skill. Experiment 3 tests how pivotality affects responsibility judgments in addition to what the action revealed about the person. Agents are held more responsible when their action was pivotal. Across experiments, our model accurately fits participants' judgments and thus provides a unified account of how people assign responsibility across a variety of situations.

3. Experiment 1: Goalies & Spinners

Participants in this experiment judged to what extent an agent was to credit for a positive outcome, or to blame for a negative one. We manipulated (i) the expectation that the chosen action will lead to success, (ii) whether the outcome was positive or negative, and (iii) the framing of the task.

The agent always faced a choice between two actions. In the *goalie world* (see Fig. 2a), participants evaluated the actions of soccer goalkeepers who could either jump to the left or to the right side, to block a shot by a striker. The outcome was positive (from the perspective of the goalkeeper) if the goalkeeper decided to jump in the direction in which the striker shot, and negative otherwise. The strikers had a tendency to shoot in one direction or another. For example, a particular striker tends to shoot to the right 80% of the time and to the left 20% of the time. The goalkeeper was said to know the striker's tendency, and also that the striker was unaware of the fact that the goalkeeper has this information. In the *spinner world* (see Fig. 2b), participants evaluated the actions of contestants in a game show whose task it was to predict the outcome of a two-colored spinner. For example, a particular contestant might face a spinner with an 80% chance of landing on yellow and 20% chance of landing on blue.

We predict that the task framing affects how people's expectations about an agent change after having observed their action, and that in turn will affect their credit and blame attributions. We predict that in the goalie world, participants will give more credit when positive outcomes resulted from *unexpected* actions, compared to expected actions. This is because in the goalie world, it is plausible that an agent could exhibit skill. Saving an unexpected ball is diagnostic for skill and positive future performance. In contrast, in the spinner world, where skill is less plausible, we predict that participants will give more credit when positive outcomes resulted from *expected* compared to unexpected actions. For negative outcomes, we predict that attributions of blame increase the more unexpected the action was for both the spinner and goalie worlds. Unexpected actions that led to negative outcomes are particularly indicative for lack of skill, and thus predictive for negative future performance.

Table 1
Pairings of decisions and outcomes used in Experiments 1 and 2. The Action column for each Player indicates the chances of that player's action being successful (assuming that the player did not have privileged access to what will actually happen). For example, in Round 1, Player 1 chose an action that had an 80% chance of being successful and Player 2 chose an action that had a 20% chance of succeeding. In fact, both agents were unsuccessful in this round as indicated by the Outcome column.

Block	Round	Action		Outcome
		Player 1	Player 2	
I	1	80%	20%	✗
I	2	80%	20%	✓
I	3	60%	40%	✗
I	4	60%	40%	✓
II	5	40%	20%	✗
II	6	40%	20%	✓
II	7	80%	60%	✗
II	8	80%	60%	✓

Note: ✗ = negative outcome, ✓ = positive outcome.

3.1. Methods

3.1.1. Participants

Participants in all experiments reported in this paper were recruited via Amazon Mechanical Turk. Only participants who live in the US, have a Human Intelligence Task (HIT) approval rate of at least 95%, and have a minimum number of 50 approved HITs were allowed to participate (see Mason & Suri, 2012, for details about Mechanical Turk). Participants were reimbursed at a rate of $6 per hour. The studies in this paper were approved by the IRB boards of University College London and Massachusetts Institute of Technology. 83 participants (39 female, $M_{\text{age}} = 34$, $SD_{\text{age}} = 11.14$) participated in this experiment. No participants were excluded from the analyses in any of our experiments. We made sure that no participant took part in more than one of our experiments.

3.1.2. Design

The framing of the task varied between participants ($N = 41$ in the *goalie world*, and $N = 42$ in the *spinner world*). We varied the *probability* that a person's action will be successful (20%, 40%, 60% and 80%), as well as whether the *outcome* was negative or positive within participants. Thus, we have a 2 (world; between) × 4 (probability of success; within) × 2 (outcome; within) design.

3.1.3. Procedure

The experiment was programmed in Flash CS5.[7] After being introduced to the basic features of the task, participants first played the game themselves for ten rounds. As goalkeepers, they decided whether to jump towards the right or left corner, and as game show contestants, they predicted which out of two colors a spinner will land on.

In each round of the main phase of the experiment, participants saw two different situations shown on the same screen (see Table 1). For example, in Round 2, Player 1's chosen action had an 80% chance of being successful whereas Player 2's action had only a 20% chance of being successful (assuming that neither player was able to anticipate what will happen). The outcome in both situations was positive, that is, both goalkeepers saved the ball in the *goalie world*, or both contestants predicted the correct outcome in the *spinner world*. Our motivation for presenting the stimuli in pairs was to highlight the contrast between situations (cf. Nagel & Waldmann, 2013). We told participants that each situation featured a different person.

We will refer to players' actions as 'expected' when the observer's expectation that the action will be successful based on the probability information is greater than 50%, and as 'unexpected' otherwise (for example, a contestant who bets that a spinner will land on a color that takes up only 20% of the spinner chose an unexpected action). Block I featured situations in which one player's action was expected and the other player's action was unexpected. In Block II, both agents' actions were either unexpected (rounds 5 and 6) or expected (rounds 7 and 8). Block I rounds were presented before Block II rounds, and the order of rounds within each block was randomized.

In each round, participants were asked to assign blame or credit, depending on the outcome. In the *goalie world*, the questions were "To what extent is each goalkeeper to blame for the goal?" or "To what extent is each goalkeeper to credit for the save?". In the *spinner world*, the questions were "To what extent is each player to blame for the negative outcome?" or "To what extent is each player to credit for the positive outcome?". Participants made their judgments on sliding scales

[7] Materials for this paper including the data, analysis files, and experiment demos can be accessed here: https://osf.io/2x3ve/.

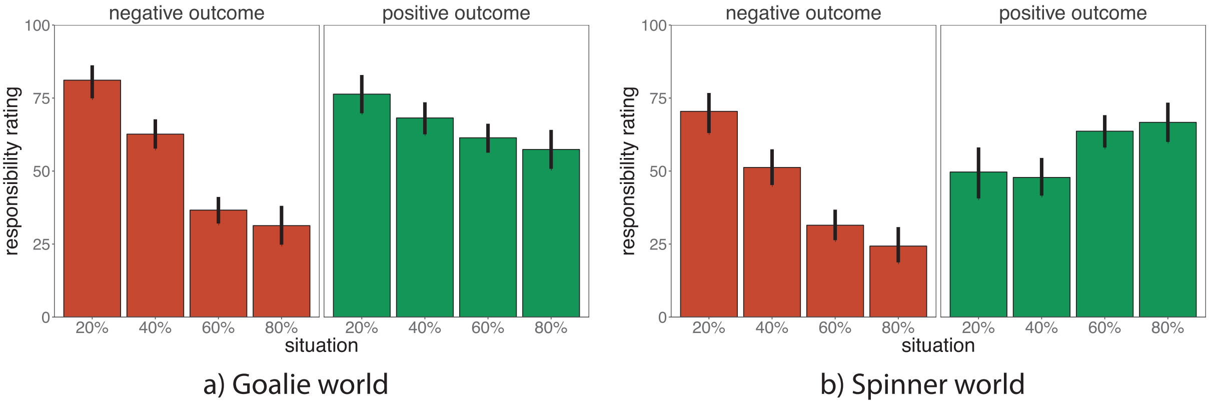


Fig. 3. Experiment 1: Mean blame (red bars) and credit (green bars) judgments in as a function of the agent's choice. *Note*: The error bars in all figures indicate 95% bootstrapped confidence intervals. The labels on the x-axis represent the expected chance that the agent's action will be successful. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

whose endpoints were labeled "not at all" (0) and "very much" (100). The slider was initiated at 0. On average, it took participants 6.19 (*SD* = 1.75) minutes to complete the experiment.

3.2. Results and discussion

In neither the *goalie world*, nor the *spinner world* were participants' responsibility judgments to a specific agent significantly influenced by what other situation was presented on the same screen.[8] For example, responsibility judgments to Player 1 in Round 1 were no different than responsibility judgments to Player 1 in Round 7. Thus, we simply aggregated participants' ratings over repeated situations. Fig. 3 shows participants' mean blame (red bars) and credit judgments (green bars) as a function of what action the agent took, separated for the goalie and spinner world. For ease of interpretation, we will discuss blame and credit judgments separately.

For *blame judgments*, there was a main effect of *world*, $F(1, 81) = 9.55$, $p = .003$, $\eta_p^2 = .11$. Blame judgments were higher in the goalie world than in the spinner world. There was also a main effect of *probability*, $F(3, 243) = 71.58$, $p < .001$, $\eta_p^2 = .47$. Participants assigned more blame the more unexpected the agent's action was. There was no interaction between *world* and *probability*, $F(3, 243) = 0.33$, $p = .806$, $\eta_p^2 < .01$.

As predicted, participants in both conditions blamed agents more for negative outcomes having resulted from unexpected compared to expected actions. A person whose action leads to a negative result reveals that they are lacking the relevant skill, which is particularly bad when they choose the unexpected course of action.

For *credit judgments*, there was a main effect of *world*, $F(1, 81) = 7.11$, $p = .009$, $\eta_p^2 = .08$. Credit judgments were higher in the goalie world than in the spinner world. There was no main effect of *probability*, $F(3, 243) = 0.55$, $p = .649$, $\eta_p^2 = .01$. However, there was an interaction between *world* and *probability*, $F(3, 243) = 7.84$, $p < .001$, $\eta_p^2 = .09$. The way in which participants assigned credit differed between conditions: whereas in the goalie world, agents received more credit for unexpected actions, in the spinner world the reverse was the case – agents were credited less when the positive outcome resulted from an unexpected action.

The interaction between action expectation and credit judgments in the goalie and the spinner worlds demonstrates that there is no direct relationship between action expectations and responsibility judgments. In Experiment 2, we will show how our model captures this interaction by assuming that participants' responsibility judgments are mediated by an inference from the observed action to a person's disposition.

4. Experiment 2: Goalies & Spinners revisited

Experiment 2 serves as both a replication of Experiment 1, and as a direct test of the proposed model. We explicitly informed participants at the beginning of the experiment that there are three different types of agents. We added a second stage to the experiment, in which participants were asked to indicate what type of agent they thought the person was after having observed their action.

In the *goalie condition*, we told participants that *average goalkeepers* base their decision to jump in one direction or another merely on the prior tendency of the striker. *Skilled goalkeepers* could correctly anticipate the striker's shot. They could save balls that were shot in the opposite direction to that of a striker's normal tendency. *Unskilled goalkeepers* did not have the ability to anticipate the striker's shot, but nevertheless thought that they did. For example, they would falsely anticipate that a striker would decide to go against their usual tendency, when in fact the striker did go with their tendency.

In the *spinner condition*, we told participants that *average players* base their decision on the probabilities of the spinner. *Skilled players* could correctly anticipate on what color the spinner would land. They could even correctly predict the outcome if the spinner landed on the less likely color. *Unskilled players* did not have the ability to anticipate on what color the spinner would land but nevertheless thought that they did. For example, they would falsely predict that a spinner would land on the less likely color, when in fact it landed on the more likely color.

Based on the results from Experiment 1, we expected to replicate the asymmetrical way in which action expectations affected attributions of blame and credit for the goalie and spinner worlds. However, we anticipated that introducing the explicit space of agents may influence people's priors to some degree. For example, most participants might a-priori think it highly unlikely that game show contestants could anticipate the outcome of a random spinner, but when told explicitly that some contestants can do this, they will update their priors accordingly. Still, we expected that participants bring their general world knowledge to bear on the task, and be more skeptical about the possibility of skill in the spinner world compared to the goalie world. Thus, we expected the differences between worlds to persist, but be smaller in Experiment 2 compared to Experiment 1.

4.1. Methods

4.1.1. Participants

82 participants ($M_{\text{age}} = 35$, $SD_{\text{age}} = 11.26$, $N_{\text{female}} = 27$) completed this experiment.

[8] For each world, we ran a within-subjects ANOVA with *probability*, *outcome*, and *context* as factors, where *context* was a dummy variable indicating the two contexts in which each situation was presented. In neither the goalie nor the spinner world, was there a main effect of *context*, or any interaction effect (all F's < 2.2 and p's > . 14).

4.1.2. Design

The design was equivalent to Experiment 1. There were $N = 41$ participants in the *goalie condition*, and $N = 41$ participants in the *spinner condition*. In addition to the *responsibility judgment phase*, the experiment also featured an *agent inference phase* in which participants judged what type of agent they thought someone was for each of the eight different situations.

4.1.3. Procedure

The initial instructions were identical to Experiment 1. Again, participants played the game themselves for ten rounds. Before completing the responsibility judgment phase, participants were instructed about the three different agent types. Blocks and trials were assigned in the same way as in Experiment 1. After completing the responsibility judgment phase, participants were told that in the second part of the experiment their task will be to judge how likely they thought that a given player was *average*, *skilled*, or *unskilled*. Participants were reminded how each agent type chose their actions. Participants were also told that they would be shown eight situations they had already seen in the first part of the experiment. For each situation, participants were instructed to distribute their belief over the three possible agent types so that the overall judgment sums up to 100%, and were given an example.

Participants indicated their judgments by typing numbers into three text boxes, one for each agent type. The order of the vertical positions in which the response options for the three agent types were presented on the screen was randomized between participants. Participants could only proceed to the next trial if their judgments added up to 100%. On average, it took participants 11.4 ($SD = 4.5$) minutes to complete the experiment.

4.2. Results

We will first report the results from the *agent inference phase*, and then look at the results from the *responsibility judgment phase*. For each phase, we first describe the model predictions, and then show how participants' empirical judgments compare.

4.2.1. Agent inferences

Model predictions. We model participants' agent inferences via Bayesian inference as described above.[9] The model predicts that participants will be more likely to believe that an agent was unskilled when the outcome was negative, and skilled when the outcome was positive. Irrespective of whether the outcome was positive or negative, participants are predicted to be more likely to believe that an agent was average for expected rather than unexpected actions. The model also predicts an interaction effect between outcome and the expected probability that an action would be successful. Whereas for negative outcomes, unexpected actions are more indicative of lack of skill, for positive outcomes, unexpected actions are more indicative of being skilled.

Empirical data. Fig. 4 shows participants' agent inferences separately for the goalie world and the spinner world. We ran ordinal logistic regressions to analyze how participants' agent inferences were affected by *world* (goalie vs. spinner), *outcome* (negative vs. positive), and the expected *probability* that an action would be successful. In order to run this analysis, we coded for each participant which agent they considered to be the most likely for each of the different situations.

There was neither a main effect of *world*, $\chi^2(1) = 0.17$, $p = .681$, nor any interaction effect involving *world*, $\chi^2(3) = 5.65$, $p = .130$. Hence, we combined the data from both worlds for further analyses. There was a main effect of *outcome*, $\chi^2(1) = 144.78$, $p < .001$. When the outcome was positive, participants were more likely to believe that the agent was skilled, and less likely to believe that the agent was unskilled, compared to situations in which the outcome was negative.

There was no main effect of *probability*, $\chi^2(1) = 0.41$, $p = .524$ but a significant interaction between *probability* and *outcome*, $\chi^2(1) = 68.18$, $p < .001$. When the outcome was negative, participants were more likely to believe that the agent was *unskilled* when he chose an unexpected action compared to an expected action. When the outcome was positive, participants were more likely to believe that the agent was *skilled* when the action was unexpected rather than expected.

Even though the results were overall not significantly affected by *world*, participants' inferences between conditions differed somewhat in situations in which an unexpected action led to a positive outcome (the two bars on the left for positive outcomes in Fig. 4). For these two cases combined, participants in the goalie world were somewhat more likely to believe that the agent was skilled (62%) and less likely to believe that he was unskilled (12%) compared to participants in the spinner world (51% skilled, and 22% unskilled). As we will see below, this difference is reflected in how participants assigned credit for unexpected actions that resulted in positive outcomes.

4.2.2. Responsibility judgments

Model predictions. In order to derive the model predictions for responsibility judgments, we need to determine the difference in expected future reward for each situation, as well as the pivotality of the agent's action for the outcome (cf. Eq. (10)). Determining pivotality here is simple: in both the goalie and the spinner world, the agent's action was always pivotal since there were only two action alternatives and the outcome would always have been different, had the agent chosen the other action.

To determine how the expectation about the agent's future reward changed, we first need to infer what participants' prior was over the different agent types in the different worlds. We infer this prior distribution $P(\mathscr{T}|\mathscr{W})$ based on participants' posterior judgments $P(\mathscr{T}|\mathscr{W}, \mathscr{S}, \mathscr{A})$ as shown in Fig. 4, assuming the agent decision functions as described in Eqs. (3)–(5).[10]

Given participants' priors and posteriors over the different agent types, we first calculate the prior expectation of future reward (Eq. (6)), and the posterior expectation of future reward (Eq. (7)), and then calculate the difference in expected reward (Eq. (8)). If the difference in expected reward is positive, the model predicts that the agent will receive credit, if the difference is negative, the model predicts blame.

Fig. 5 shows the model predictions for the goalie and spinner worlds, where we combined the 'Difference in expected reward' and 'Pivotality for the outcome' according to Eq. (10). For both the goalie and spinner world, the model predicts that agents will be blamed more for negative outcomes when they took unexpected actions. In the goalie world, the model predicts that agents will receive more credit for positive outcomes that resulted from unexpected actions. In the spinner world, the model predicts no strong differences in credit as a function of how expected the agent's action was. The model's different predictions for how credit will be assigned in the goalie and spinner world are a result of using participants' agent inferences to calculate the difference in expected future reward. Compared to participants in the goalie condition, participants in the spinner condition were more likely to believe that the agent may be unskilled for positive outcomes that resulted from unexpected actions, and somewhat less likely to believe that the agent was skilled when the positive outcome resulted from an expected action.

Empirical data. The red and green bars in Fig. 5 show participants' mean blame and credit judgments separately for the goalie and the spinner worlds.[11]

[9] See Appendix A for a detailed worked example of how the model makes agent inferences, causal attributions, and assigns responsibility.

[10] Table A1 in the Appendix shows what priors best explain participants' agent inferences across the three experiments reported in this paper.

[11] Like in Experiment 1, there was no effect of the context in which a situation was presented (all F's < 2.2 and p's $> .09$).

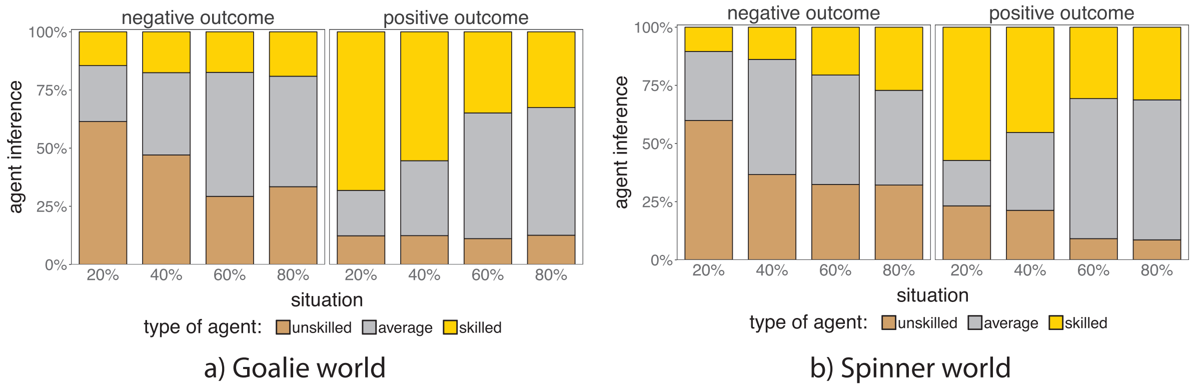


Fig. 4. Mean agent inferences in (a) the goalie world, and (b) the spinner world, as a function of what action the agent chose (x-axis), and whether the outcome was negative or positive (panels). For example, in the goalie world, when the goalkeeper chose to jump in the 20% direction and didn't save the ball (leftmost bar), the participants were most likely to believe that the agent was unskilled, less likely to believe that he was average, and least likely to believe that the agent was skilled.

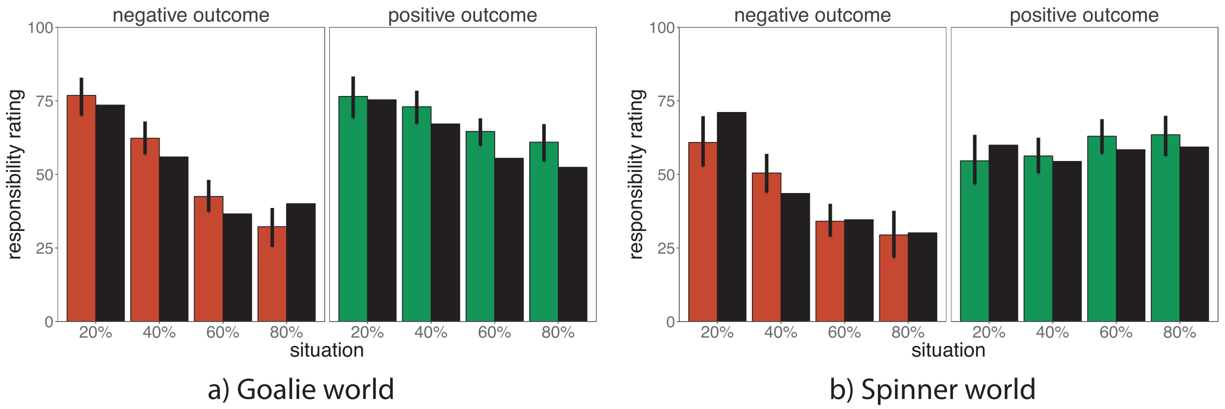


Fig. 5. Experiment 2: Mean responsibility ratings (colored bars) and model predictions (black bars) in the (a) goalie world, and (b) spinner world. Error bars indicate bootstrapped 95% confidence intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For *blame judgments*, there was a main effect of *world*, $F(1, 80) = 9.52$, $p = .003$, $\eta_p^2 = .11$. Blame judgments were higher in the goalie world than in the spinner world. There was also a main effect of *probability*, $F(3, 240) = 44.53$, $p < .001$, $\eta_p^2 = .36$. Participants assigned more blame the lower the chance that the agent's action would succeed. There was no interaction between *world* and *probability*, $F(3, 240) = 1.16$, $p = .325$, $\eta_p^2 = .01$.

For *credit judgments*, there was a main effect of *world*, $F(1, 80) = 7.91$, $p = .006$, $\eta_p^2 = .09$. Credit judgments were higher in the goalie world than in the spinner world. There was no main effect of *probability*, $F(3, 240) = 0.23$, $p = .873$, $\eta_p^2 < .01$. However, there was an interaction between *world* and *probability*, $F(3, 240) = 3.98$, $p = .009$, $\eta_p^2 = .05$. Whereas for the goalie world, participants assigned more credit when actions had a low probability of succeeding, in the spinner world, credit judgments increased slightly with the probability that the action would be successful.

4.3. Discussion

Overall, the results of Experiment 2 closely replicate what we found in Experiment 1. As Fig. 5 shows, the model's predictions are closely in line with participants' judgments. The model correctly predicts that agents' blame for negative outcomes increases, the more unexpected their action was. Note that in the goalie world, the model actually predicts that a goalie who jumped in the 80% corner will be blamed more than a goalie who jumped in the 60% corner. The model makes this prediction because participants were somewhat more likely to infer that the goalie was unskilled in the 80% case compared to the 60% case (cf. Fig. 4a). For positive outcomes, the model correctly captures how credit attributions in the goalie world decreased the more expected the goalie's action was. In the spinner world, the model captures that participants' credit attributions were mostly flat across the range of probability values. We compare the quantitative fit of our model with that of other models in Section 6 ("Model comparison across experiments").

Remember that unlike in Experiment 1, Experiment 2 explicitly informed participants about the three different agent types at the beginning of the experiment. The fact that the pattern of results for the goalie condition looked almost identical in Experiment 1 and Experiment 2 shows that instructing participants about the different agent types did not affect their responsibility judgments. However, in the spinner condition, introducing the different agent types did affect participants' judgments – particularly judgments of credit for positive outcomes. Instead of giving more credit for expected compared to unexpected actions, averaged credit attributions did not show an effect of action expectation anymore.

We investigated whether this overall pattern of results might be due to an aggregation effect whereby some participants gave more credit to unexpected over expected actions and vice versa for another group of participants. Indeed, a clustering analysis of individual participants' responsibility judgments revealed that there were two subgroups of participants in both the goalie and the spinner condition. The subgroups did not differ in how they assigned blame. However, they did differ in how action expectations affected their credit judgments.[12]

[12] See Appendix B for a detailed analysis of individual differences in Experiment 2.

While in the goalie condition, the subgroup which gave more credit for unexpected actions was larger than the subgroup who assigned more credit for expected actions, in the spinner condition, the subgroups were about equal size thus leading to the overall flat pattern across action expectations. Importantly, there was a very close correspondence between participants' responsibility judgments and agent inferences in each subgroup. Participants who believed that an unexpected action was diagnostic of a skilled agent, gave more credit for unexpected compared to expected positive outcomes. Participants who were more reluctant to believe in skill for these cases, gave more credit for expected compared to unexpected positive outcomes.

What these results demonstrate is that our introduction of the different agent types did not change participants' conception of the game for the goalie world very much. They naturally considered skill to be an important component of the game. In the spinner world, many participants did not consider it plausible that an unexpected positive outcome could have resulted from some special skill. Only when explicitly informing participants about this possibility, did a subgroup deem unexpected positive outcomes more creditworthy than expected ones.

The results of Experiments 1 and 2 support the idea that participants' responsibility judgments are influenced by what an action reveals about a person, and how participants' expectations about the person's future behavior compare with their prior expectations. So far, however, we have not manipulated the causal role that the person's action played in bringing about the outcome. In general, we believe that a person's responsibility for the outcome increases the more prominently their action featured in bringing it about. A variety of aspects influence the perceived causal role of a person's action. In Experiment 3, we manipulated whether or not the agent's action was pivotal for bringing about the outcome.

5. Experiment 3: Pivotal Gardeners

The scenario in Experiment 3 was inspired by a scenario featured in Johnson and Rips (2013). A gardener must choose between two fertilizers to make a flower grow. One fertilizer generally has a better chance of making the flower grow. The gardener decides to use one of the two fertilizers, and the flower either grows or does not grow. In addition to what actually happened, we informed participants about what *would* have happened if the gardener had chosen differently. Fig. 6 shows the eight different situations that participants saw in the experiment. For example, in Fig. 6a, the gardener chose the fertilizer that generally had a lower chance of making the flower grow. In this situation, the fertilizer failed to make the flower grow. However, the flower would have failed to grow even if the gardener had chosen the alternative fertilizer that had a higher chance of success. Because the outcome would have been the same no matter what the gardener did, his action was not pivotal in this situation. In Fig. 6b, the gardener's action is the same – he chooses the low-probability fertilizer and the flower doesn't grow. However, this time, the alternative fertilizer would have made the flower grow. If participants' responsibility judgments are affected by pivotality, then we would expect participants to blame the gardener in Fig. 6b more for the negative outcome than the gardener in Fig. 6a.

More generally, we predict that in addition to participants' expectations about how well a gardener is likely to do in the future, responsibility judgments will be affected by whether the gardener's action was pivotal. Gardeners whose action was pivotal are predicted to be held more responsible than gardeners whose action was not pivotal.

5.1. Methods

5.1.1. Participants

41 participants ($M_{\text{age}} = 34$, $SD_{\text{age}} = 10.61$, $N_{\text{female}} = 17$) completed this experiment.

5.1.2. Design

We manipulated the gardener's *decision* (expected vs. unexpected), the *pivotality* (pivotal vs. not pivotal) of the gardener's decision, and the resulting *outcome* (negative vs. positive). All factors were manipulated within participants. As in Experiment 2, participants first made responsibility judgments, and then agent inferences.

5.1.3. Procedure

Participants were first introduced to the scenario. Their task was to evaluate gardeners who decided which of two fertilizers to use in order to make a flower grow. One of the fertilizers had a 30% chance of making the flower grow, and the other one had a 70% chance (cf. Fig. 6). After seeing several examples, participants were introduced to the three agent types, who differed in their ability to sense whether the different fertilizers are actually going to make a particular flower grow.

Average gardeners do not have the skill to sense whether or not a particular fertilizer will make the flower grow and rely on the probability with which the different fertilizers have made similar flowers grow in the past, as indicated on the labels on the bottles. *Skilled gardeners* do have the skill to sense whether or not a particular fertilizer will make the flower grow in a particular case. Thus, they might sense that a fertilizer with a low growth probability will nonetheless make the flower grow in this particular case, or that a fertilizer with a high growth probability will fail to make the flower grow in this particular case. *Bad gardeners* do not have the skill to sense whether or not a particular fertilizer will make the flower grow. However, they falsely believe that they do. Thus, they might sense that a fertilizer with a low growth probability will actually make the flower grow in this particular case when in fact it will not.

Participants then had to answer a number of comprehension check questions, and they were only allowed to proceed with the experiment if they answered all questions correctly. If they answered one or more questions incorrectly, they were asked to carefully reread the instructions.

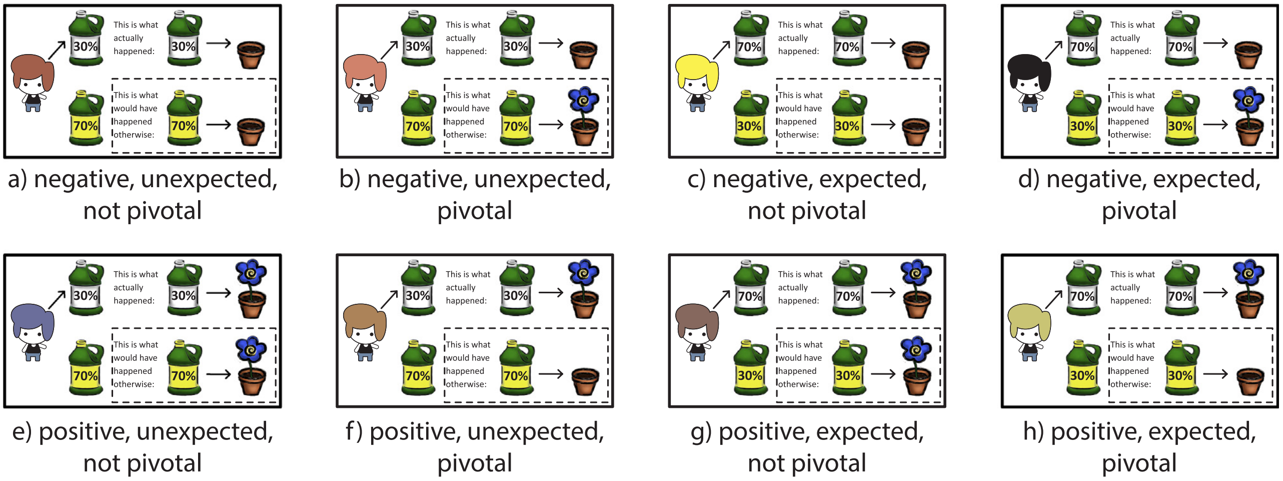
In the responsibility judgment phase, participants saw the negative (Fig. 6a-d) and positive situations (Fig. 6e-h) on the screen in two sets of four. We counterbalanced which set was presented first. The four situations were presented simultaneously on the screen in a 2 × 2 grid. The position of a situation in the grid was randomized. We presented the situations simultaneously to highlight the differences between them. Participants were asked to say to what extent each of the gardeners was to blame/credit for the negative/positive outcome. Participants indicated their responses on sliders as in Experiment 2.

After participants finished the responsibility judgment phase, they were instructed that their next task will be to judge how likely a given gardener was *average*, *skilled*, or *bad* for each of the different situations they had seen in the first part of the experiment. They were reminded how the different gardener types made their decisions. In the agent inference phase, participants were shown the eight different situations in a randomized order. Agent inference judgments were elicited in the same way as in Experiment 2. It took participants 10.4 ($SD = 4.9$) minutes on average to complete the experiment.

5.2. Results

5.2.1. Agent inferences

Model predictions. For the situations in which the gardener's action was pivotal, the predictions of the model are essentially identical to the predictions in Experiment 2 (taking into account the probabilities of the different actions being successful). For situations in which the gardener's action was not pivotal, the model predicts that participants will be just as likely to infer that the agent was skilled or bad irrespective of the actual outcome. This is the case because these agents choose actions based on the true reward that an action will yield. In non-pivotal cases, the rewards for each action are the same – either both fertilizers make the flower grow, or they both don't. However, which action the agent


This is what actually happened:
This is what would have happened otherwise:
30%
70%

a) negative, unexpected, not pivotal

b) negative, unexpected, pivotal

c) negative, expected, not pivotal

d) negative, expected, pivotal

e) positive, unexpected, not pivotal

f) positive, unexpected, pivotal

g) positive, expected, not pivotal

h) positive, expected, pivotal

Fig. 6. Different situations in Experiment 3. The situations differ in whether the outcome was positive or negative, whether the agent's action was unexpected or expected, and in whether the agent's action was pivotal or not pivotal.

took in non-pivotal cases is still diagnostic for whether he is average.

Empirical data. Fig. 7a shows participants' mean agent inferences for the eight different situations. An ordinal logistic regression revealed that there was a main effect of *outcome*, $\chi^2(1) = 89.23$, $p < .001$. Participants were more likely to believe that the agent was skilled and less likely to believe that he was bad for positive compared to negative outcomes. There was neither a main effect of *decision*, $\chi^2(1) = 0.67$, $p = .414$, nor of *pivotality*, $\chi^2(1) = 0.13$, $p = .724$. However, there was a interaction between *outcome* and *decision*, $\chi^2(1) = 68.10$, $p < .001$. For negative outcomes, participants were less likely to believe that the agent was bad when their action was expected compared to unexpected. For positive outcomes, participants were less likely to believe that the agent was skilled for expected compared to unexpected actions.

There was also an interaction between *outcome* and *pivotality*, $\chi^2(1) = 4.29$, $p = .039$. For negative outcomes, participants were more likely to believe that the agent was bad when their action was pivotal rather than non-pivotal. For positive outcomes, participants were more likely to believe that the agent was skilled for pivotal versus non-pivotal actions. Finally, there was a three-way interaction between *outcome*, *decision*, and *pivotality*, $\chi^2(1) = 4.79$, $p = .029$.

Overall, these results are largely consistent with the model predictions. However, participants' posterior inferences also deviated from the model's predictions in systematic ways. For example, when the outcome was negative and the agent's action not pivotal, participants were more likely to believe that the agent was bad rather than skilled. Similarly, when the outcome was positive and the action non-pivotal, participants were more likely to believe that the agent was skilled rather than bad. This suggests that participants' inferences about the agents may have been richer than what we capture with our agent decision models. For example, participants may have assumed that skilled gardeners generally make flowers bloom, and bad gardeners generally fail at making flowers bloom. That is, even if two gardeners are given exactly the same two fertilizers, a skilled gardener will make it work, whereas a bad gardener is likely to fail.

5.2.2. Responsibility judgments

Model predictions. We calculated the difference in expected future reward in the same way as we did for Experiment 2. Whereas in Experiment 2 the agent's action was always pivotal, here the pivotality of an agent's action was reduced for some cases. We model these cases by assuming that one change would have been required to render the agent's action pivotal (the alternative outcome would have needed to be different from what it actually was). Since one change would have been required to make the agent's action pivotal, their pivotality for the outcome in these cases was $\frac{1}{2}$ (cf. Eq. (9)).

Fig. 7b shows the model predictions in black. The model predicts that gardeners will be held more responsible for pivotal actions than for non-pivotal actions. It further predicts that agents will be held more responsible when they had chosen an unexpected action. Finally, the model predicts an interaction effect. The differences in responsibility for pivotal versus non-pivotal actions are predicted to be greater when the actions were unexpected.

Empirical data. The red and green bars in Fig. 7b show participants'

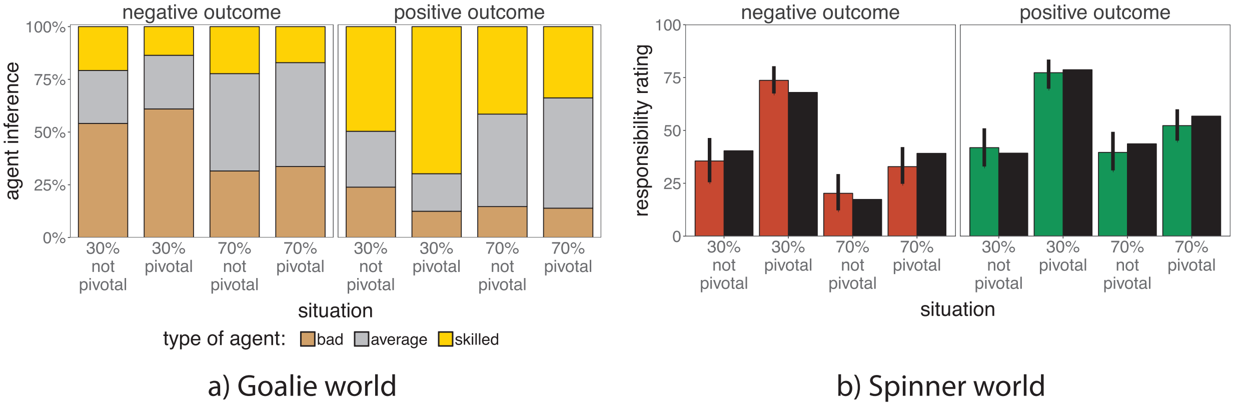


Fig. 7. Experiment 3: Mean agent inferences (a) and responsibility judgments (b) in the gardener experiment as a function of what action the agent chose, and whether the outcome was negative or positive. Black bars in (b) show model predictions. Error bars indicate bootstrapped 95% confidence intervals.

blame and credit judgments, respectively.

For blame judgments, there was a main effect of *decision*, $F(1, 40) = 45.29$, $p < .001$, $\eta_p^2 = 0.53$. Gardeners were blamed more when their decision was unexpected than when it was expected. There was a main effect of *pivotality*, $F(1, 40) = 32.19$, $p < .001$, $\eta_p^2 = 0.45$. Gardeners were blamed more when their decision was pivotal than when it was not. There was also an interaction between *decision* and *pivotality*, $F(1, 40) = 19.91$, $p < .001$, $\eta_p^2 = 0.33$. Blame judgments increased more strongly with pivotality for unexpected decisions than for expected decisions.

Similarly, for credit judgments, there was a main effect of *decision*, $F(1, 40) = 7.26$, $p = .010$, $\eta_p^2 = 0.15$. Gardeners received more credit when their decision was unexpected than when it was expected. There was a main effect of *pivotality*, $F(1, 40) = 31.97$, $p < .001$, $\eta_p^2 = 0.44$. Credit judgments were greater for pivotal compared to non-pivotal decisions. Finally, there was an interaction effect between *decision* and *pivotality*, $F(1, 40) = 15.28$, $p = .0003$, $\eta_p^2 = 0.28$. Credit increased more as a function of pivotality for unexpected compared to expected decisions.

5.3. Discussion

As in Experiment 2, there was a close correspondence between the predictions of our model and participants' responsibility judgments. Both dispositional inferences about the person, as well as the causal role that the action played in bringing about the outcome influenced participants' judgments. For example, even though participants' agent inferences were very similar in situations in which the agent's action was unexpected and didn't work out (the first two bars in Fig. 7a), participants assigned significantly more blame when the agent's action was pivotal. More generally, agents received the most blame when their action was indicative of poor judgment, and when what they did mattered in the actual situation. Conversely, agents received the most credit when their action was indicative of skill, and their action was pivotal.

6. Model comparison across experiments

So far, we have discussed the correspondence between model predictions and participants' judgments only qualitatively. We will now look at how well the model does in quantitative terms, and compare the predictions of our model to that of other models.

Fig. 8 shows the combined results for Experiment 2 and 3 for different versions of our model.[13] Across these experiments, a model that includes both 'difference in expectation' as well as 'pivotality' as predictors (Fig. 8a) accounts very well for participants' responsibility judgments. Table 2 compares the different models on a number of goodness-of-fit measures.

Likelihood ratio tests reveal that the full model does significantly better than a model which only considers 'difference in expectation' as a predictor $\chi^2(1) = 54.3$, $p < .001$ (Fig. 8b), or a model which only considers 'pivotality' as predictor $\chi^2(1) = 44.1$, $p < .001$ (Fig. 8c). In line with Johnson and Rips' (2015) optimality account, we also fitted a model to the data that makes its predictions based on whether the agent's action was optimal from the observer's perspective, and whether the outcome was positive or negative. We coded actions as optimal when the agent went with the option that was most likely to lead to a positive outcome from the observer's perspective. Even though the optimality model has the same number of free parameters as our full model, it cannot account for participants' responsibility judgments as well (Fig. 8d).[14]

The model comparisons show that both components of our model are required to adequately capture participants' responsibility judgments across the range of experiments we considered in this paper. The fact that the optimality model fails to capture participants' judgments, shows that there is no direct link from action expectations to responsibility judgments. How much credit or blame a person receives depends on how our posterior expectations about a player differ from our prior expectations, based on having observed their action.[15]

7. General discussion

How do people hold others responsible for the outcomes of their actions? In this paper, we have shown that the answer to this question depends on at least the following two cognitive processes: (i) a dispositional inference about what an action reveals about a person, and how this compares to our prior expectations, as well as (ii) a causal inference that considers what role a person's action played in bringing about the outcome. We developed a computational account of responsibility attribution that models both of these processes and quantitatively compared the predictions of the model to people's responsibility attributions across a set of experiments.

In order to model the inference from a person's action to their disposition, we need a theory of mind that captures the way in which agents make decisions (cf. Fig. 1). Dispositional inferences can then be modeled via inverting the decision-making process, using Bayesian inference to update the prior belief over what kind of person the agent is. This dispositional inference matters since it affects our expectations about the person's behavior in the future. We find that, in line with the predictions of our model, people assigned more credit for positive outcomes to the extent that they believed the action was indicative of skill. Conversely, they assigned more blame for negative outcomes when they believed the action was indicative of bad decision-making.

In order to model the causal inference that connects a person's action with the outcome, we need a causal model of the situation that allows us to assess the extent to which the person's action made a difference. We use a modified counterfactual model of actual causation (Chockler & Halpern, 2004; Halpern & Pearl, 2005; Lagnado et al., 2013) to define how close a person's action was to having made a difference to the outcome. Participants assigned more responsibility when the person's action made a difference in the actual situation, and less responsibility when the action would only have mattered if things had turned out differently.

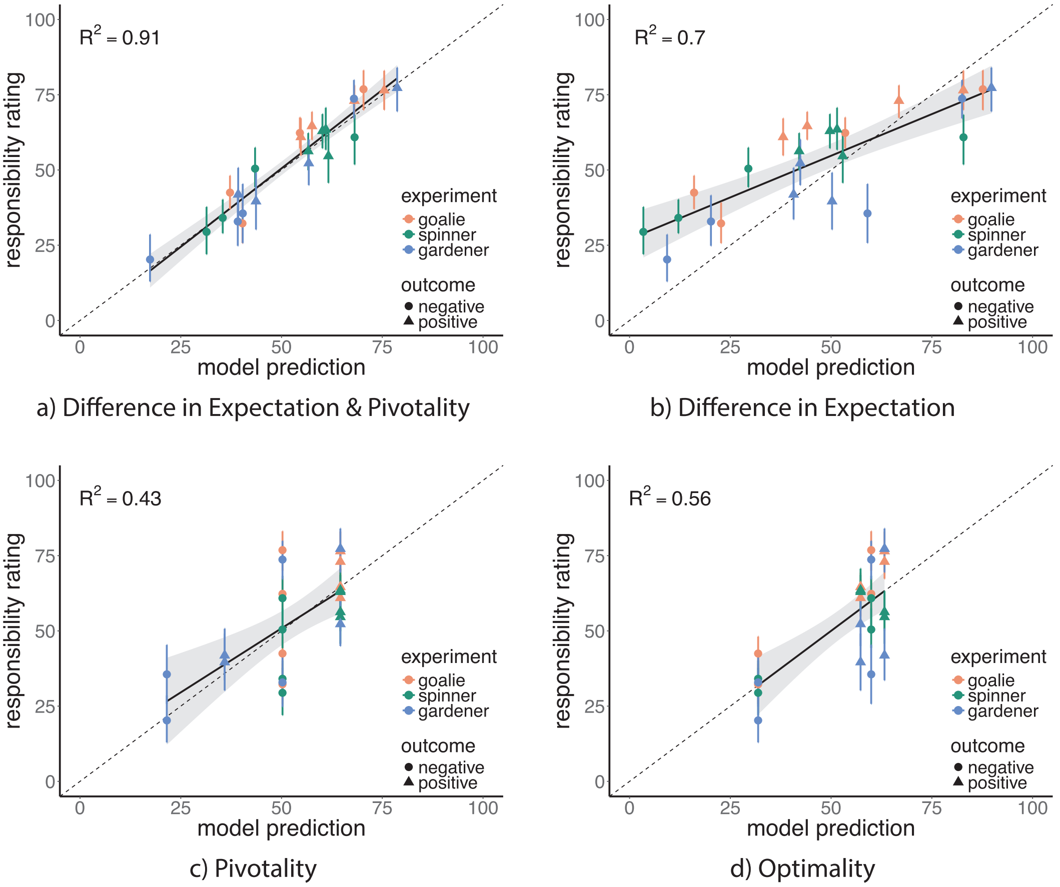
Our model solves a puzzle in the literature about the relationship between action expectations and judgments of responsibility. Most theories predict that someone will be held more responsible for an outcome when her action was unexpected (Brewer, 1977; Fincham & Jaspars, 1983; Petrocelli et al., 2011; Spellman, 1997). However, as Johnson and Rips (2015) have shown, this does not always hold true. Sometimes people are held more responsible for actions that they were expected to take. Instead of going directly from action expectations to judgments of responsibility, our model shows how this process is mediated by a dispositional inference. While doing the right thing is often the same as doing the expected thing (Johnson & Rips, 2015), our experiments tease these two factors apart. The result is intuitive: we blame others who did not meet our expectations when things go wrong, and we credit others whose actions went beyond our expectations when things go right.

Beyond the empirical contributions of this paper, we provide a unified framework that explains how people assign responsibility to others making decisions under uncertainty. Our model quantitatively captures participants' responsibility judgments in a variety of different situations. While previous work considered the importance of

[13] Experiment 1 is not included because participants did not make agent inferences in this experiment, and so we do not have direct model predictions.

[14] Because agents are predicted to be blamed more when their action was suboptimal, and credited more when their action was optimal, we included an interaction term in the regression. Thus, our implementation of the optimality model is the following: $Responsibility = \alpha_0 + \alpha_1 optimal + \alpha_2 outcome + \alpha_3(optimal \times outcome)$.

[15] See Tables C1 and C2 in the Appendix for an overview of the model's predictions across all experiments and situations.



a) Difference in Expectation & Pivotality

b) Difference in Expectation

c) Pivotality

d) Optimality

Fig. 8. Scatterplots of model predictions (x-axis) and mean responsibility ratings (y-axis) across all three experiments. The black lines show the best-fitting regression line for each model together with the 95% confidence interval shown in gray. Error bars indicate bootstrapped 95% confidence intervals.

Table 2
Summary of the model results. Values for *r* and RMSE indicate means (with 5% and 95% quantiles in parentheses) based on 100 split-half cross-validation runs. BIC scores are based on running the models on the full data set.

Model	*r*	RMSE	BIC
Difference & pivotality	.86 (.66, .95)	10.56 (6.17, 17.21)	158.59
Difference	.70 (.30, .90)	26.92 (16.4, 40.6)	209.74
Pivotality	.63 (.41, .77)	14.23 (11.39, 17.54)	199.53
Optimality	.66 (.42, .84)	14.55 (10.54, 17.91)	199.47

Note: BIC = Bayesian Information Criterion (lower values indicate better model performance).

dispositional inferences and causal attribution for how people assign responsibility, our model is the first to translate these ideas into quantitative predictions. By linking dispositional inferences to changes in expectations about future behavior, and causal attributions to the pivotality of a person's action, we provide concrete proposals that can be tested in a fine-grained manner and expanded on by future research.

While we believe that our model captures some key aspects of how people assign responsibility, there is a still a long way to go to arrive at a comprehensive computational account of responsibility attribution. In the following, we compare the model presented here with previous accounts of responsibility attribution. We then highlight two aspects where we believe the model raises interesting questions for future research. The first question concerns the role that expectations play in our model. The second question concerns extending the model presented here to capture people's judgments of *moral* responsibility.

7.1. Relationship to previous accounts of responsibility attribution

Responsibility is a rich and multi-faceted concept (Fishbein & Ajzen, 1973; Gailey & Falk, 2008; Hart, 2008; Heider, 1958; Shaver & Drown, 1986; Vincent, 2011). There are many factors that influence the way in which people assign responsibility to an agent for an outcome, such as what the agent's role or obligation was in the situation (Hamilton, 1978, 1980; Malle, 2004), how much control the agent had over her action and the outcome (Ajzen, 1971; Alicke, 2000; Wells & Gavanski, 1989; Young & Phillips, 2011), whether the agent was able to foresee the consequences of her actions (Lagnado & Channon, 2008; Markman & Tetlock, 2000), whether the consequences were intended (Cushman, 2008; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015; Shultz & Wright, 1985), whether the consequences were realized in the intended way (Alicke & Rose, 2012; Guglielmo & Malle, 2010; Pizarro, Uhlmann, & Bloom, 2003), and how bad (or good) the consequences turned out to be (Robbennolt, 2000; Schroeder & Linder, 1976).

There have been several attempts to integrate these different factors into coherent theoretical frameworks (Alicke, 2000; Malle et al., 2014; Schlenker, Britt, Pennington, Murphy, & Doherty, 1994; Shaver, 1985; Weiner, 1995). For example, Malle et al.'s (2014) path model of blame postulates a sequence of tests that an observer goes through when assigning blame for a negative outcome, such as whether the agent was causally responsible for the outcome, whether the action violated a norm, whether the agent intended the outcome, and so on. However, as with most previous theories of responsibility attribution, the model does not make any quantitative predictions. The model we introduced in this paper only considers a subset of the factors identified in the literature, but expresses these factors in formally precise terms. This allows our model to generate quantitative predictions that capture the way in which participants' assign responsibility across a range of different settings.

Counterfactual thinking plays a key part in our model (cf. Brewer, 1977; Byrne, 2016; Gerstenberg & Lagnado, 2012; Kahneman & Miller, 1986; Lipe, 1991; Petrocelli et al., 2011; Spellman, 1997). Counterfactuals are also central for how causation is analyzed in the law (Hart & Honoré, 1959/1985; Lagnado & Gerstenberg, 2017; Schaffer, 2005; Schaffer, 2010; Stapleton, 2008). Tthe "but-for" test aims to establish whether the agent's action was a cause of the outcome by asking if a different outcome would have come about but for the agent's action. The "reasonable person" test compares a person's actions with how a reasonable person would have behaved in the same situation (Green, 1967; Lloyd-Bostock, 1979). Our model incorporates both of these ideas: We use an extended but-for test to express the causal role that the person's action played in bringing about the outcome (cf. Chockler & Halpern, 2004). And we use an analogue of the reasonable person test whereby we compare an observer's posterior expectations about a person after having observed their action, with our prior expectations (cf. Fincham & Jaspars, 1983; Petrocelli et al., 2011). These prior expectations capture the observer's belief about how a reasonable person would have behaved in the actual situation, as well as similar situations.

7.2. The role of expectations

Expectations play a central part in our model, but our experiments did not directly address the question of where initial expectations come from. There are at least two possible options here: Our expectations might be based on what *any reasonable person* would do, or on what *the specific person* would do. Because we have a model that translates differences in expectations to judgments of responsibility, we can dig deeper and ask what source of expectations accounts best for participants judgments. To illustrate, imagine that Michael Jordan, one of the best basketball players of all time, ends up scoring 18 points in a game and his team loses. To what extent do we blame Jordan for the team's loss? For most players, scoring 18 points is excellent. However, Jordan scored on average around 30 points. So, the extent to which we blame Jordan depends on what we adopt as the comparison standard to compute our prior expectations.

If we consider a person-specific norm, then Jordan should receive a lot of blame, since we expected for him to do better. However, if we consider a population-specific norm and take into account how well players generally do, then Jordan's performance still compares well (cf. Kelley, 1973; Sytsma et al., 2012). Our model provides a computational framework for incorporating these different kinds of norms, and for testing how different norms of comparison influence responsibility judgments.

The existence of multiple norms as reference points also opens the door for motivated attribution effects (Alicke, 2000). An observer who likes Jordan may highlight the fact that Jordan's performance was still better than how others would be expected to perform if they were in his shoes. An observer who dislikes Jordan may choose the person-specific norm as a reference point, and put emphasis on the fact that Jordan failed to deliver this time.

7.3. Richer agent models and moral responsibility

In this paper, we focused on how people assign responsibility to decision-makers for outcomes in achievement contexts (Frieze & Weiner, 1971; Gerstenberg, Ejova, & Lagnado, 2011; Weiner, 1985; Weiner et al., 1972). While the agent models we developed capture a range of phenomena, they are relatively simple in that they only vary along the dimension of skill. Richer agent models are required to apply our framework to other contexts that may raise questions of moral responsibility (cf. Gerstenberg et al., 2015; Parkinson & Byrne, 2017). When judging moral responsibility, different aspects about a person's character become relevant. In moral evaluations, we care not only about other people's skill, but also about them being fair, honest, trustworthy, empathetic, etc. (Allen et al., 2015; Kleiman-Weiner, Shaw, & Tenenbaum, 2017; Lapsley & Lasky, 2001; Walker & Hennig, 2004).

In line with our model's emphasis on the role of dispositional inferences for judging responsibility, recent work in moral psychology (see Waldmann, Nagel, & Wiegmann, 2012, for a review) has argued for a shift from an action-centered approach toward a person-centered approach to moral judgment (Bartels & Pizarro, 2011; Bayles, 1982; Pizarro & Tannenbaum, 2011; Pizarro, Uhlmann, & Salovey, 2003; Uhlmann et al., 2015; Uhlmann & Zhu, 2013; Uhlmann, Zhu, & Tannenbaum, 2013). For example, Uhlmann et al. (2015, p. 72) argue that "current act-based theories in moral psychology provide an incomplete account of moral judgment to the extent that they do not include the fundamental human motivation to determine the moral character of others" (see also Lagnado & Gerstenberg, 2017). The person-centered perspective puts the focus on a central social function of moral evaluations (Rai & Fiske, 2011; Scanlon, 2009): figuring out who is good and who is bad.

While Uhlmann et al. (2015) discuss formal models for inferring a person's character from their actions (Ditto & Jemmott, 1989; Fiske, 1980; McKenzie & Mikkelsen, 2007; Nelson, 2005), there is no formal model to date which translates this character inference into a moral judgment (although see Kleiman-Weiner et al., 2015, for a computational account on how inferences about an agent's intentions influence moral judgments). Our model takes a first step in that direction by suggesting that character inferences affect moral judgments through changed expectations about future behavior.

8. Conclusion

People have a rich understanding of how the world works, and how other people work (Gerstenberg & Tenenbaum, 2017; Wellman & Gelman, 1992). With this understanding come expectations of how events will unfold. Expectations are a key construct in the cognitive sciences ranging from the predictive coding theory about the brain (Kilner, Friston, & Frith, 2007) to what determines our moment-to-moment happiness (Rutledge, Skandali, Dayan, & Dolan, 2014). In this paper, we argue that expectations also play a key role for how we hold others responsible. We credit others when our expectations about how they will act in the future are better than our prior expectations, and blame them if our expectations are lower. It is important to maintain accurate models of other people through updating our expectations, as we constantly have to decide who to interact with, and who to avoid (Baumard, André, & Sperber, 2013; Rai & Fiske, 2011). By combining a person-centered viewpoint on how a person's action changed our expectations about their future behavior, with an action-centered viewpoint on the causal role that the person's action played in bringing about the outcome, we arrive at a more comprehensive understanding of how we hold others responsible.

Acknowledgments

We thank Kelsey Allen for extensive comments on the manuscript, and Kevin Smith for valuable feedback and discussion. This work was supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216 and by an ONR grant N00014-13-1-0333. The results of Experiment 1 have previously been reported in Gerstenberg, Ullman, Kleiman-Weiner, Lagnado, and Tenenbaum (2014).

Appendix A. Detailed example of how the model works

We illustrate how the model works by applying it to participants' agent inferences and responsibility judgments in Experiment 2. Participants are predicted to update their prior beliefs about what agent type a person is, to a posterior belief after having observed the person's action using Bayesian inference (cf. Eq. (1)). The decision-making functions describe the likelihood with which each of the different agent type takes the observed action (cf. Eqs. (3)–(5)). Because we did not ask participants for their prior beliefs, we inferred their priors from their posterior judgments (see Table A1).[16] Comparing model predictions to participants' agent inferences thus allows us to gage the extent to which our formal agent functions correspond to participants' beliefs about how the agents make decisions based on the informal descriptions we had given them.

To determine the difference in expectations, we first have to calculate the expected reward for each agent type by integrating over the possible situations that could arise in each world. Let us illustrate by example, using the goalie world. To determine the probability distribution over different situations $P(S|w = \text{goalie})$ we considered hypothetical strikers with priors ranging from 0.1 to 0.9 to shoot towards the left corner, and crossed this with whether they actually shot towards the left or the right corner. Each prior for the striker was considered equally likely, whereby the strikers' actual shots adhered to their prior tendency (i.e. a striker who had an 80% tendency to shoot the ball towards the left direction also did so with $p = 0.8$). This means that a situation in which a striker shot towards the 80% direction was more likely to happen than a situation in which the same striker shot towards the 20% direction.[17]

The expected reward for the average goalie, for example, can then be calculated as

Table A1
The inferred priors $P(\mathscr{T}|\mathscr{W})$ over the three agent types for the three different worlds considered in this paper.

	Unskilled agent	Average agent	Skilled agent
Goalie world	0.26	0.43	0.31
Spinner world	0.27	0.46	0.27
Gardener world	0.30	0.37	0.33

$$\mathbb{E}[r|\mathscr{W} = \text{goalie},\mathscr{T} = \text{average}] = \sum_{a\in\mathscr{A}}\sum_{s\in\mathscr{S}} r_a\cdot P(a|s,\ \mathscr{T} = \text{average})\cdot P(s|\mathscr{W} = \text{goalie}), \tag{11}$$

where we use the fitted decision noise parameter β that is part of the agent's decision-making function $P(a|s,\ \mathscr{T} = \text{average})$ (cf. Eq. (3)). Repeating the same procedure for unskilled, average, and skilled goalies results in expected rewards of 0.19, 0.58, and 0.81, respectively. This means that an unskilled goalie would be expected to save 19% of the shots if it faced the whole range of strikers that are part of the possible situations we considered. A skilled goalie would be expected to save 81% of the shots. Note that the expected future rewards for each agent type are identical in the spinner world since the two worlds are structurally isomorphic. To facilitate comparison between different experiments, we rescaled the expected rewards such that the unskilled agent has an expected reward of 0, and the skilled agent has an expected reward of 1, which resulted in an expected reward of 0.63 for the average agent.

Now that we know the expected reward of each agent type $\mathbb{E}[r|\mathscr{W} = w,\ \mathscr{T} = t]$, and the priors over the different agent types $P(t|W = w)$, we can determine the prior expected reward for each world by integrating over the different agent types as shown in Eq. (6). For example, in the *goalie world*, we get

$$\begin{aligned}\mathbb{E}[r|\mathscr{W} = \text{goalie}] = {} & P(t = \text{unskilled}|w)\cdot\mathbb{E}[r|w,\ t = \text{unskilled}]+\\ & P(t = \text{average}|w)\cdot\mathbb{E}[r|w,\ t = \text{average}]+\\ & P(t = \text{skilled}|w)\cdot\mathbb{E}[r|w,\ t = \text{skilled}]\\ = {} & 0.26\cdot 0 + 0.43\cdot 0.63 + 0.31\cdot 1\\ = {} & 0.58,\end{aligned} \tag{12}$$

where we used the rescaled expected future rewards for each agent type. We perform the same calculations for the spinner world.

We determine the posterior expected reward for each situation according to Eq. (7). For $P(T = t|\mathscr{W} = w,\ \mathscr{A} = a_{\text{obs}},\ \mathscr{S} = s_{\text{obs}})$, we took participants' mean agent inferences as shown in Fig. 4. For example, consider the situation in the goalie world in which the goalkeeper decided to jump in the 20% direction and did not save the ball (i.e. the leftmost bar of the empirical data in Fig. 4a). In this case, participants' mean judgments indicate that they thought there was a 61% chance that the agent was unskilled, a 24% chance the agent was average, and a 15% chance the agent was skilled. Accordingly, the posterior expected reward we get based on having observed the agent jumping in the 20% direction and failing to save the ball is

$$\begin{aligned}\mathbb{E}[r|\mathscr{W} = \text{goalie},a_{obs},\ s_{obs}] = {} & P(t = \text{unskilled}|w,\ a_{obs},\ s_{obs})\cdot\mathbb{E}[r|w,\ t = \text{unskilled}]+\\ & P(t = \text{average}|w,\ a_{obs},\ s_{obs})\cdot\mathbb{E}[r|w,\ t = \text{average}]+\\ & P(t = \text{skilled}|w,\ a_{obs},\ s_{obs})\cdot\mathbb{E}[r|w,\ t = \text{skilled}]\\ & = 0.61\cdot 0 + 0.24\cdot 0.63 + 0.15\cdot 1\\ & = 0.30.\end{aligned} \tag{13}$$

We can then calculate the difference in expected reward by subtracting the prior expected reward $\mathbb{E}[r|\mathscr{W} = \text{goalie}]$ from the posterior expected reward $\mathbb{E}[r|\mathscr{W} = \text{goalie},a_{obs},\ s_{obs}]$ as shown in Eq. (8). For this particular example, the model's expectation about the agent's future reward has decreased from a prior of 0.58 to a posterior of 0.30. We perform the same calculations for the remaining situations that arose in the experiment. In

[16] To infer the prior, we calculated $P(\mathscr{T}|\mathscr{W}) \propto \frac{P(\mathscr{T}\mid\mathscr{W},\mathscr{S},\mathscr{A})}{P(\mathscr{A}\mid\mathscr{S},\mathscr{T})}$ for each situation and then fit $P(\mathscr{T}|\mathscr{W})$ using least squares.

[17] Note that this was not the case in our experiment, due to the balanced design we had chosen. For example, across the 8 trials that participants saw, a striker with a 20% to shoot in the left direction was just as likely to shoot in the left direction as a striker with an 80% tendency to shoot left.

order to predict participants' responsibility judgments, we then take the difference in expectation for each of the situations, as well as the agent's pivotality, and fit the weighting of both components as shown in Eq. (10).[18] Table C1 shows participants' agent inferences as well as the model predictions for all situations in Experiments 2 and 3, and Table C2 shows participants' responsibility judgments and model predictions.

Appendix B. Analysis of individual differences

To investigate the extent to which the overall pattern of responsibility judgments in the goalie and spinner worlds of Experiment 2 was due to an aggregation of systematically different groups of participants (see Fig. 5), we clustered participants based on their responsibility judgments using a Gaussian Mixture Model (Fraley & Raftery, 2002; Fraley, Raftery, Murphy, & Scrucca, 2012). For both the spinner and the goalie world, the best solution was found by assuming two clusters of participants.[19]

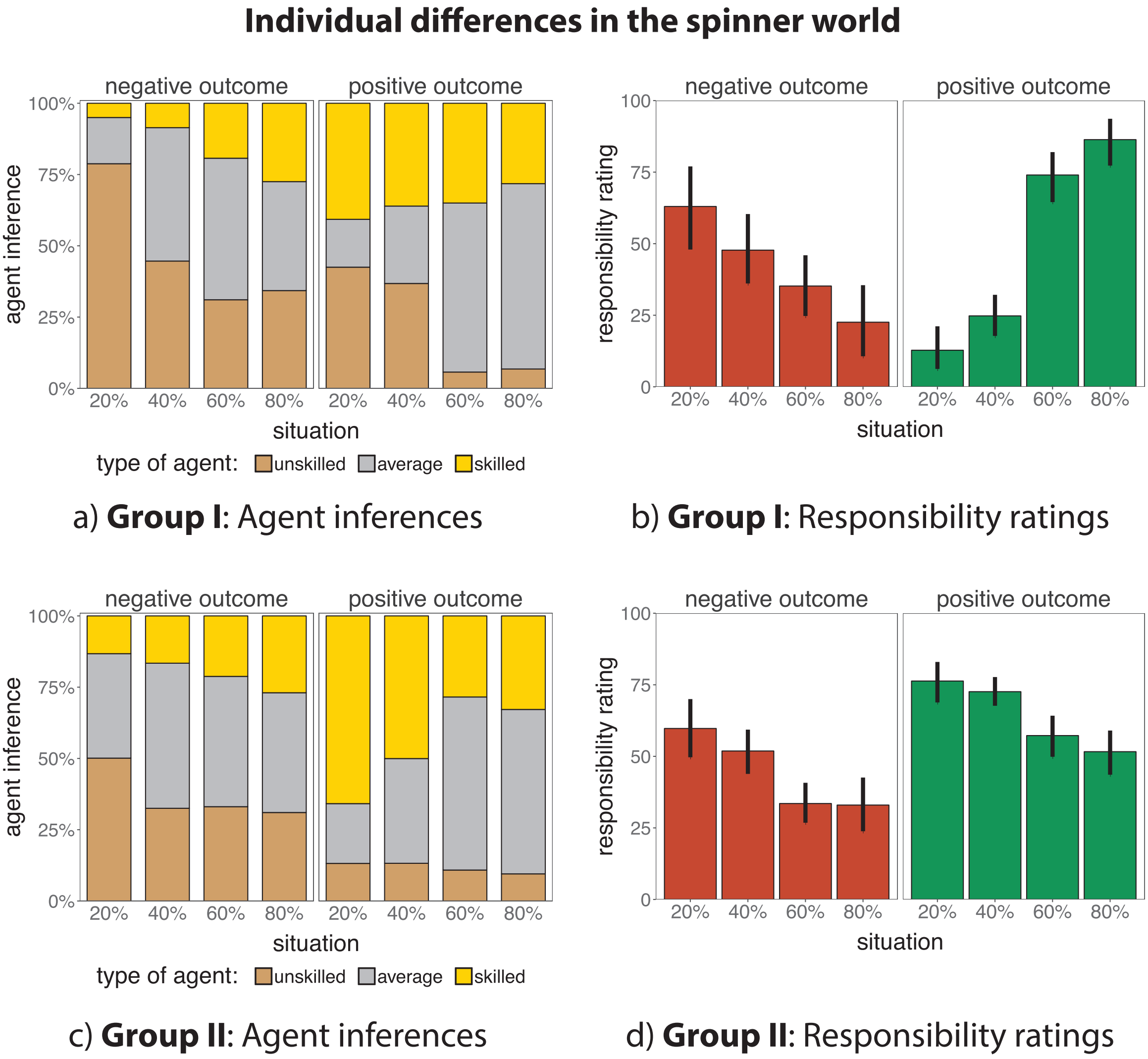
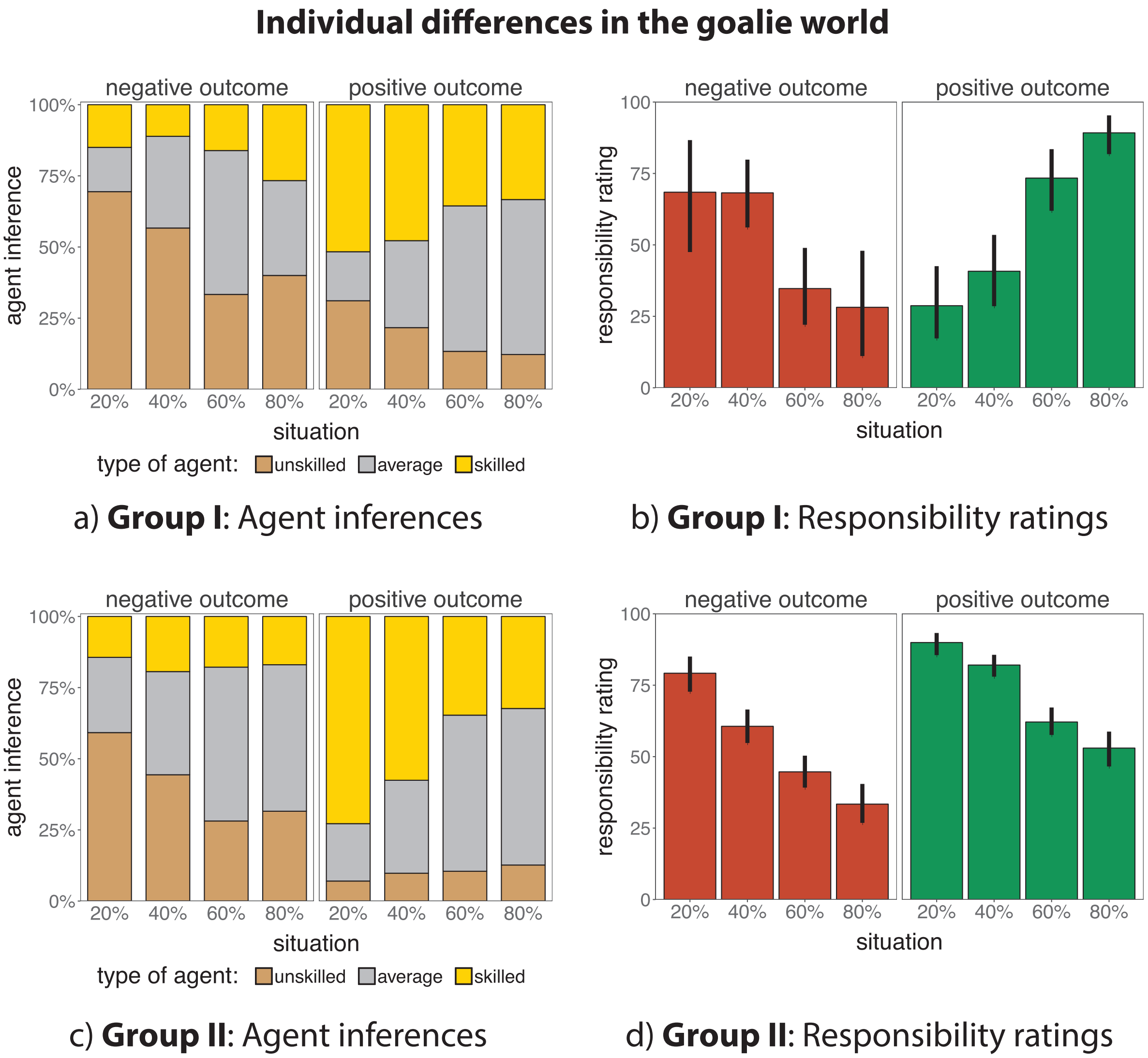


Fig. B1. Experiment 2, Spinner condition: Agent inferences and responsibility ratings for different clusters of participants in the *spinner world*. Group I ($N = 14$) gives more credit for expected successes while Group II ($N = 27$) gives more credit for unexpected successes. Error bars indicate bootstrapped 95% confidence intervals.

Fig. B1 shows the agent inferences on the left, and the responsibility judgments on the right for the two identified clusters of participants in the spinner condition. For negative outcomes, both groups drew similar agent inferences and blame judgments. However, for positive outcomes, the two groups differed markedly. The two groups' agent inferences differed most strongly for situations in which the positive outcome resulted from an unexpected action. In these situations, Group I ($N = 14$) considered it just as likely that the agent may be unskilled or skilled. Group II ($N = 27$), in contrast, thought that it was much more likely that the agent was skilled than unskilled in these situations. In line with the predictions of our model, there was a close correspondence between the inferences participants drew about the players and how they assigned responsibility. Whereas Group I

[18] Note that while in the spinner and goalie experiment, the pivotality term is equal to a dummy variable for outcome (since the agent's action is always pivotal), the results of Experiment 3 show that participants are indeed sensitive to the pivotality of a person's action rather than just the outcome.

[19] The BIC scores for 1, 2, 3, or 4 clusters were 3034, 3002, 3016, 3021, for the goalie world, and 3154, 3130, 3141, 3152 for the spinner world where lower values indicate better clustering solutions. BIC scores continued to get worse for larger cluster numbers.



a) **Group I**: Agent inferences

b) **Group I**: Responsibility ratings

c) **Group II**: Agent inferences

d) **Group II**: Responsibility ratings

Fig. B2. Experiment 2, Goalie condition: Agent inferences and responsibility ratings for different clusters of participants in the *goalie world*. Group I ($N = 9$) gives more credit for expected successes while Group II ($N = 32$) gives more credit for unexpected successes. Error bars indicate bootstrapped 95% confidence intervals.

assigned more credit when the agent's action was expected, Group II assigned more credit for positive outcomes that resulted from unexpected actions.

One way to understand this pattern of results is that Group I contains participants who were skeptical of the possibility that someone could exhibit skill in games of chance. Even though we told participants that some of the players were able to foresee what color the spinner is going to land on, participants in this group seem to have remained suspicious when positive outcomes resulted from unexpected actions (cf. Hilton, Fein, & Miller, 1993). Participants in this group perhaps did not suspend their belief that in games of chance the optimal thing to do is to bet on the most likely outcome.

Participants in Group II were inclined to believe that correctly anticipating an unexpected outcome was indicative of skill (either a-priori or after reading our instructions about the existence of such agents) and thus assigned more credit for unexpected than for expected actions.

The fact that there was overall no influence of the probability manipulation on the credit judgments (cf. Fig. 5) resulted from combining these two groups of participants who differed in what inferences they drew from observing an agent's unexpected action resulting in a positive outcome. Our model – which predicts participants' responsibility judgments in terms of how the observer's posterior expectations about the player compare with the prior expectations – captures the different ways in which the two groups assign responsibility.

Fig. B2 shows the results of running the same clustering analysis for participants in the goalie world. Again, we found two clusters of participants that differed both in their agent inferences for the unexpected positive outcomes and in how they assigned responsibility for these cases. Participants in Group I ($N = 9$) believed that there was a good chance that the goalie may have been unskilled when the positive outcome had resulted from an unexpected action. Accordingly, they assigned less credit in these cases compared to when the goalie had saved a ball that was shot in the expected direction. In contrast, participants from Group II ($N = 32$) appeared to reason that the goalie must have been skilled when he saved a ball that was shot in the unexpected direction. Participants in this group gave most credit when the goalie saved unexpected shots.

So, both in the spinner and the goalie world we found two groups of participants that differed in their prior assumptions about the plausibility of skill, which in turn influenced their inferences. One group of participants considered skill to be an unlikely factor and accordingly attributed less credit for unexpected than for expected successes. The other group considered skill more likely and thus gave *more* credit for unexpected than for expected successes. Note that while both types of groups existed in both the spinner and goalie condition of the experiment, the proportion of participants falling in each group differed between conditions in a predictable manner. In the spinner world, the number of participants who were suspicious about unexpected positive outcomes being due to skill was greater (35%) compared to the goalie world (22%).

Appendix C. Agent inferences and responsibility judgments across all experiments.

See Tables C1 and C2.

Table C1
Agent inference phase: Model predictions and empirical means for each experiment and situation. *Note*: The **prob** column shows which action the agent took, the **out** column shows whether the outcome was negative (0) or positive (1), the **predicted posterior** and **empirical posterior** columns show the models' predicted agent inferences and participants' mean agent inferences for the bad, average, and skilled agent, respectively.

Exp	Prob	Out	Pivotality	Predicted posterior			Empirical posterior		
				Bad	Average	Skilled	Bad	Average	Skilled
Goalie	20	0	Pivotal	53.53	31.63	14.85	61.44	24.05	14.51
Goalie	40	0	Pivotal	46.87	40.13	13.00	47.07	35.37	17.56
Goalie	60	0	Pivotal	41.29	47.25	11.45	29.27	53.29	17.44
Goalie	80	0	Pivotal	37.21	52.46	10.32	33.41	47.49	19.10
Goalie	20	1	Pivotal	11.69	29.43	58.88	12.32	19.51	68.17
Goalie	40	1	Pivotal	10.32	37.67	52.00	12.39	32.20	55.41
Goalie	60	1	Pivotal	9.16	44.68	46.15	11.10	54.02	34.88
Goalie	80	1	Pivotal	8.30	49.88	41.82	12.56	54.88	32.56
Spinner	20	0	Pivotal	53.90	33.46	12.63	59.90	29.66	10.44
Spinner	40	0	Pivotal	46.86	42.16	10.98	36.66	49.49	13.85
Spinner	60	0	Pivotal	41.04	49.34	9.62	32.39	47.05	20.56
Spinner	80	0	Pivotal	36.82	54.55	8.63	32.15	40.71	27.15
Spinner	20	1	Pivotal	12.66	33.48	53.86	23.20	19.54	57.27
Spinner	40	1	Pivotal	11.00	42.18	46.82	21.27	33.46	45.27
Spinner	60	1	Pivotal	9.63	49.36	41.00	9.10	60.22	30.68
Spinner	80	1	Pivotal	8.64	54.57	36.79	8.59	60.17	31.24
Gardener	30	0	Non-pivotal	33.36	30.05	36.60	54.10	25.12	20.78
Gardener	30	0	Pivotal	55.14	30.66	14.20	60.98	25.44	13.59
Gardener	70	0	Non-pivotal	26.99	43.41	29.61	31.56	46.15	22.29
Gardener	70	0	Pivotal	44.44	44.12	11.44	33.66	49.27	17.07
Gardener	30	1	Non-pivotal	33.36	30.05	36.60	23.88	26.49	49.63
Gardener	30	1	Pivotal	12.43	29.46	58.11	12.44	17.80	69.76
Gardener	70	1	Non-pivotal	26.99	43.41	29.61	14.66	43.85	41.49
Gardener	70	1	Pivotal	10.10	42.71	47.19	13.90	52.32	33.78

Table C2
Responsibility judgment phase: Model predictions and empirical means for each experiment and situation. *Note*: The **probability** column shows which action the agent took, the **outcome** column shows whether the outcome was negative (0) or positive (1), the **difference** column shows the model's predicted difference in expected future reward after having observed the action and outcome, the pivotality column shows whether or not the agent's action was pivotal in the situation, the **prediction** column shows the model's predicted responsibility (taking into account both difference and pivotality), and the **rating** column shows participants' mean responsibility ratings. Note that the table shows the unnormalized values for **difference** and **pivotality**.

Experiment	Probability	Outcome	Difference	Pivotality	Prediction	Rating
Goalie	20	0	−0.28	−1.00	−70.35	−76.84
Goalie	40	0	−0.18	−1.00	−54.55	−62.29
Goalie	60	0	−0.07	−1.00	−37.21	−42.50
Goalie	80	0	−0.09	−1.00	−40.31	−32.24
Goalie	20	1	0.23	1.00	75.47	76.50
Goalie	40	1	0.18	1.00	68.08	73.00
Goalie	60	1	0.11	1.00	57.55	64.60
Goalie	80	1	0.09	1.00	54.79	60.95
Spinner	20	0	−0.27	−1.00	−68.13	−60.85
Spinner	40	0	−0.11	−1.00	−43.44	−50.46
Spinner	60	0	−0.06	−1.00	−35.41	−34.10
Spinner	80	0	−0.03	−1.00	−31.39	−29.43
Spinner	20	1	0.14	1.00	61.63	54.62
Spinner	40	1	0.10	1.00	56.63	56.27
Spinner	60	1	0.13	1.00	60.16	62.98
Spinner	80	1	0.13	1.00	60.98	63.49
Gardener	30	0	−0.20	−0.50	−40.37	−35.56
Gardener	30	0	−0.27	−1.00	−67.97	−73.68
Gardener	70	0	−0.05	−0.50	−17.37	−20.27
Gardener	70	0	−0.08	−1.00	−39.15	−32.88
Gardener	30	1	0.10	0.50	39.25	41.85
Gardener	30	1	0.25	1.00	78.69	77.29
Gardener	70	1	0.13	0.50	43.66	39.61
Gardener	70	1	0.10	1.00	56.76	52.27

References

Ajzen, I. (1971). Attribution of dispositions to an actor: Effects of perceived decision freedom and behavioral utilities. *Journal of Personality and Social Psychology, 18*(2), 144–156.

Ajzen, I., & Fishbein, M. (1975). A Bayesian analysis of attribution processes. *Psychological Bulletin, 82*(2), 261–277.

Ajzen, I., & Fishbein, M. (1978). Use and misuse of Baye's theorem in causal attribution: Don't attribute it to Ajzen and Fishbein either. *Psychological Bulletin, 85*(2), 244–246.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*(4), 556–574.

Alicke, M. D., & Rose, D. (2012). Culpable control and causal deviance. *Journal of Personality and Social Psychology Compass, 6*, 723–725.

Allen, K., Jara-Ettinger, J., Gerstenberg, T., Kleiman-Weiner, M., & Tenenbaum, J. B. (2015). Go fishing! responsibility judgments when cooperation breaks down. In D. C. Noelle, (Ed.). *Proceedings of the 37th annual conference of the Cognitive Science Society* (pp. 84–89). Austin, TX: Cognitive Science Society.

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 doi:10.1038%2Fs41562-017-0064.

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

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition, 121*(1), 154–161. http://dx.doi.org/10.1016/j.cognition.2011.05.010.

Baumard, N., André, J. B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences, 36*(01), 59–78. http://dx.doi.org/10.1017/s0140525x11002202.

Bayles, M. D. (1982). Character, purpose, and criminal responsibility. *Law and Philosophy, 1*(1), 5–20.

Botti, S., & McGill, A. L. (2006). When choosing is not deciding: The effect of perceived responsibility on satisfaction. *Journal of Consumer Research, 33*(2), 211–219. http://dx.doi.org/10.1086/506302.

Brewer, M. B. (1977). An information-processing approach to attribution of responsibility. *Journal of Experimental Social Psychology, 13*(1), 58–69.

Byrne, R. M. (2016). Counterfactual thought. *Annual Review of Psychology, 67*, 135–157.

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research, 22*(1), 93–115.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380.

Dennett, D. C. (1987). *The intentional stance.* Cambridge, MA: MIT Press.

Ditto, P. H., & Jemmott, J. B. (1989). From rarity to evaluative extremity: Effects of prevalence information on evaluations of positive and negative characteristics. *Journal of Personality and Social Psychology, 57*(1), 16–26. http://dx.doi.org/10.1037/0022-3514.57.1.16.

Duff, R. A. (1993). Choice, character, and criminal liability. *Law and Philosophy, 12*(4), 345–383. http://dx.doi.org/10.1007/bf01000637.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin, 99*(1), 3–19.

Fincham, F. D., & Jaspars, J. M. (1983). A subjective probability approach to responsibility attribution. *British Journal of Social Psychology, 22*(2), 145–161.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a bayesian perspective. *Psychological Review, 90*(3), 239–260.

Fischhoff, B., & Lichtenstein, S. (1978). Don't attribute this to reverend Bayes. *Psychological Bulletin, 85*(2), 239–243.

Fishbein, M., & Ajzen, I. (1973). Attribution of responsibility: A theoretical note. *Journal of Experimental Social Psychology, 9*(2), 148–153.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology, 38*(6), 889–906. http://dx.doi.org/10.1037/0022-3514.38.6.889.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association, 97*, 611–631.

Fraley, C., Raftery, A. E., Murphy, T. B., & Scrucca, L. (2012). *mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation*. Technical Report No. 597.

Frieze, I., & Weiner, B. (1971). Cue utilization and attributional judgments for success and failure. *Journal of Personality, 39*(4), 591–605.

Gailey, J. A., & Falk, R. F. (2008). Attribution of responsibility as a multidimensional concept. *Sociological Spectrum, 28*(6), 659–680.

Gerstenberg, T., Ejova, A., & Lagnado, D. A. (2011). Blame the skilled. In C. Carlson, C. Hölscher, & T. Shipley (Eds.). *Proceedings of the 33rd annual conference of the Cognitive Science Society* (pp. 720–725). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.). *Proceedings of the 34th annual conference of the Cognitive Science Society* (pp. 378–383). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2014). From counterfactual simulation to causal judgment. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.). *Proceedings of the 36th annual conference of the Cognitive Science Society* (pp. 523–528). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle, (Ed.). *Proceedings of the 37th annual conference of the Cognitive Science Society* (pp. 782–787). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Halpern, J. Y., & Tenenbaum, J. B. (2015). Responsibility judgments in voting scenarios. In D. C. Noelle, (Ed.). *Proceedings of the 37th annual conference of the Cognitive Science Society* (pp. 788–793). Austin, TX: Cognitive Science Society.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition, 115*(1), 166–171.

Gerstenberg, T., & Lagnado, D. A. (2012). When contributions make a difference: Explaining order effects in responsibility attributions. *Psychonomic Bulletin & Review, 19*(4), 729–736.

Gerstenberg, T., & Lagnado, D. A. (2014). Attributing responsibility: Actual and counterfactual worlds. In J. Knobe, T. Lombrozo, & S. Nichols (Vol. Eds.), *Oxford studies in experimental philosophy: Vol. 1*, (pp. 91–130). Oxford University Press.

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science, 28*(12), 1731–1744 doi:10.1177%2F0956797617713053.

Gerstenberg, T., & Tenenbaum, J. B. (2016). Understanding "almost": Empirical and computational studies of near misses. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.). *Proceedings of the 38th annual conference of the Cognitive Science Society* (pp. 2777–2782). Austin, TX: Cognitive Science Society.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldmannn (Ed.). *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

Gerstenberg, T., Ullman, T. D., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2014). Wins above replacement: Responsibility attributions as counterfactual replacements. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.). *Proceedings of the 36th annual conference of the Cognitive Science Society* (pp. 2263–2268). Austin, TX: Cognitive Science Society.

Gilbert, D. T. (1998). Ordinary personology. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.). *Handbook of social psychology* (pp. 89–150). (4th ed.). New York: McGraw-Hill.

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., ... Tenenbaum, J. B. (2006). Intuitive theories of mind: A rational approach to false belief. In R. Sun, & N. Miyake (Eds.). *Proceedings of the 28th annual conference of the Cognitive Science Society* (pp. 1382–1387). Austin, TX: Cognitive Science Society.

Green, E. (1967). The reasonable man: Legal fiction or psychosocial reality? *Law & Society Review, 2*, 241–258.

Guglielmo, S., & Malle, B. F. (2010). Enough skill to kill: Intentionality judgments and the moral valence of action. *Cognition, 117*(2), 139–150.

Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *British Journal for the Philosophy of Science, 66*, 413–457.

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.

Hamilton, V. L. (1978). Who is responsible? Toward a social psychology of responsibility attribution. *Social Psychology, 41*(4), 316–328.

Hamilton, V. L. (1980). Intuitive psychologist or intuitive lawyer? Alternative models of the attribution process. *Journal of Personality and Social Psychology, 39*(5), 767–772.

Hart, H. L. A. (2008). *Punishment and responsibility.* Oxford: Oxford University Press.

Hart, H. L. A., & Honoré, T. (1959/1985). *Causation in the law.* New York: Oxford University Press.

Heider, F. (1958). *The psychology of interpersonal relations.* John Wiley & Sons Inc.

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

Hilton, J. L., Fein, S., & Miller, D. T. (1993). Suspicion and dispositional inference. *Personality and Social Psychology Bulletin, 19*(5), 501–512. http://dx.doi.org/10.1177/0146167293195003.

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

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*(10), 785 doi:10.1016%2Fj.tics.2016.08.007.

Johnson, S. G., & Rips, L. J. (2015). Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive Psychology, 77*, 42–76. http://dx.doi.org/10.1016/j.cogpsych.2015.01.003.

Johnson, S. G. B., & Rips, L. J. (2013). Good decisions, good causes: Optimality as a constraint on attribution of causal responsibility. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.). *Proceedings of the 35th annual conference of the Cognitive Science Society* (pp. 2662–2667). Austin, TX: Cognitive Science Society.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*(2), 136–153.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist, 28*(2), 107–128.

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing, 8*(3), 159–166.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In D. C. Noelle, (Ed.). *Proceedings of the 37th annual conference of the Cognitive Science Society* (pp. 1123–1128). Austin, TX: Cognitive Science Society.

Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. B. (2017). Constructing social preferences from anticipated judgments: When impartial inequity is fair and why? In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.). *Proceedings of the 39th annual conference of the Cognitive Science Society* (pp. 676–681). Austin, TX: Cognitive Science Society.

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

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

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition, 108*(3), 754–770.

Lagnado, D. A., & Gerstenberg, T. (2015). A difference-making framework for intuitive judgments of responsibility. In D. Shoemaker (Vol. Ed.), *Oxford studies in agency and responsibility: Vol. 3*, (pp. 213–241). Oxford University Press.

Lagnado, D. A., & Gerstenberg, T. (2017). Causation in legal and moral reasoning. In M. Waldmann (Ed.). *Oxford handbook of causal reasoning* (pp. 565–602). Oxford University Press.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science, 47*, 1036–1073.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. In A. Gopnik, & L. Schulz (Eds.). *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). Oxford University Press.

Lapsley, D. K., & Lasky, B. (2001). Prototypic moral character. *Identity, 1*(4), 345–363. http://dx.doi.org/10.1207/s1532706xid0104.

Leonhardt, J. M., Keller, L. R., & Pechmann, C. (2011). Avoiding the risk of responsibility by seeking uncertainty: Responsibility aversion and preference for indirect agency when choosing for others. *Journal of Consumer Psychology, 21*(4), 405–441. http://dx.doi.org/10.1016/j.jcps.2011.01.001.

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

Lipe, M. G. (1991). Counterfactual reasoning as a framework for attribution theories. *Psychological Bulletin, 109*(3), 456–471.

Livengood, J. (2011). Actual causation and simple voting scenarios. *Noûs*, 1–33.

Lloyd-Bostock, S. (1979). The ordinary man, and the psychology of attributing causes and responsibility. *The Modern Law Review, 42*(2), 143–168. http://dx.doi.org/10.1111/j.1468-2230.1979.tb01522.x.

Luce, R. (1959). *Individual choice behavior: A theoretical analysis.* John Wiley.

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction.* Cambridge, MA: MIT Press.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186. http://dx.doi.org/10.1080/1047840x.2014.877340.

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

Markman, K. D., & Tetlock, P. E. (2000). 'i couldn't have known': Accountability, foreseeability and counterfactual denials of responsibility. *British Journal of Social Psychology, 39*(3), 313–325. http://dx.doi.org/10.1348/014466600164499.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*(1), 1–23.

McKenzie, C. R., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology, 54*(1), 33–61.

Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review, 102*(2), 331–355.

Morse, S. J. (2003). Diminished rationality, diminished responsibility. *Ohio State Journal of Criminal Law, 1*, 289–308.

Nagel, J., & Waldmann, M. R. (2013). Deconfounding distance effects in judgments of moral obligation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Nelson, J. D. (2005). Finding useful questions: On bayesian diagnosticity, probability, impact, and information gain. *Psychological Review, 112*(4), 979–999. http://dx.doi.org/10.1037/0033-295x.112.4.979.

Nordbye, G. H. H., & Teigen, K. H. (2014). Being responsible versus acting responsibly: Effects of agency and risk taking on responsibility judgments. *Scandinavian journal of psychology, 55*(2), 102–114.

Pantelis, P. C., Baker, C. L., Cholewiak, S. A., Sanik, K., Weinstein, A., Wu, C. C., ... Feldman, J. (2014). Inferring the intentional states of autonomous virtual agents. *Cognition, 130*(3), 360–379. http://dx.doi.org/10.1016/j.cognition.2013.11.011.

Parkinson, M., & Byrne, R. M. J. (2017). Moral judgments of risky choices: A moral echoing effect. *Judgment and Decision Making, 12*(3), 236–252.

Pearl, J. (2000). *Causality: Models, reasoning and inference.* Cambridge, England: Cambridge University Press.

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

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer, & P. R. Shaver (Eds.). *The social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). Washington, DC: APA Press.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology, 39*(6), 653–660.

Pizarro, D. A., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science, 14*(3), 267–272. Retrieved from < http://www.ncbi.nlm.nih.gov/pubmed/12741752 >.

Rachlinski, J. (2002–2003). Misunderstanding ability, misallocating responsibility. *Brooklyn Law Review, 68*, 1055–1091.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review, 118*(1), 57–75. http://dx.doi.org/10.1037/a0021867.

Robbennolt, J. K. (2000). Outcome severity and judgments of "responsibility": A meta-analytic review. *Journal of Applied Social Psychology, 30*(12), 2575–2609.

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences, 111*(33), 12252–12257. http://dx.doi.org/10.1073/pnas.1407535111.

Scanlon, T. M. (2009). *Moral dimensions.* Harvard University Press.

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

Schaffer, J. (2010). Contrastive causation in the law. *Legal Theory, 16*(04), 259–297.

Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., & Doherty, K. (1994). The triangle model of responsibility. *Psychological Review, 101*(4), 632–652.

Schroeder, D. A., & Linder, D. E. (1976). Effects of actor's causal role, outcome severity, and knowledge of prior accidents upon attributions of responsibility. *Journal of Experimental Social Psychology, 12*(4), 340–356.

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

Shaver, K. G., & Drown, D. (1986). On causality, responsibility, and self-blame: A theoretical note. *Journal of Personality and Social Psychology, 50*(4), 697–702.

Shultz, T. R., & Wright, K. (1985). Concepts of negligence and intention in the assignment of moral responsibility. *Canadian Journal of Behavioural Science, 17*(2), 97–108.

Sinnott-Armstrong, W., & Levy, K. (2011). Insanity defenses. In J. D. D. Dolinko (Ed.). *Oxford handbook on the philosophy of the criminal law*. Oxford University Press.

Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives.* USA: Oxford University Press.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology, 66*(1), 223–247. http://dx.doi.org/10.1146/annurev-psych-010814-015135.

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General, 126*(4), 323–348.

Stapleton, J. (2008). Choosing what we mean by 'causation' in the law. *Missouri Law Review, 73*(2), 433–480.

Stephan, S., Willemsen, P., & Gerstenberg, T. (2017). Marbles in inaction: Counterfactual simulation and causation by omission. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.). *Proceedings of the 39th annual conference of the Cognitive Science Society* (pp. 1132–1137). Austin, TX: Cognitive Science Society.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction.* Cambridge University Press.

Sytsma, J., Livengood, J., & Rose, D. (2012). Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(4), 814–820.

Trope, Y. (1974). Inferential processes in the forced compliance situation: A Bayesian analysis. *Journal of Experimental Social Psychology, 10*(1), 1–16.

Trope, Y., & Burnstein, E. (1975). Processing the information contained in another's behavior. *Journal of Experimental Social Psychology, 11*(5), 439–458.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10*(1), 72–81.

Uhlmann, E. L., & Zhu, L. L. (2013). Acts, persons, and intuitions: Person-centered cues and gut reactions to harmless transgressions. *Social Psychological and Personality Science, 5*(3), 279–285. http://dx.doi.org/10.1177/1948550613497238.

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition, 126*(2), 326–334. http://dx.doi.org/10.1016/j.cognition.2012.10.005.

van Inwagen, P. (1978). Ability and responsibility. *The Philosophical Review, 87*(2), 201–224. http://dx.doi.org/10.2307/2184752.

Vincent, N. A. (2011). A structured taxonomy of responsibility concepts. In N.a. Vincent, I. van de Poel, & J. van den Hoven (Eds.). *Moral responsibility: Beyond free will and determinism.* Dordrecht: Springer.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. *The oxford handbook of thinking and reasoning* (pp. 364–389). New York: Oxford University Press.

Walker, L. J., & Hennig, K. H. (2004). Differing conceptions of moral exemplarity: Just, brave, and caring. *Journal of Personality and Social Psychology, 86*(4), 629–647. http://dx.doi.org/10.1037/0022-3514.86.4.629.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92*(4), 548.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct.* New York: The Guilford Press.

Weiner, B., Heckhausen, H., Meyer, W. U., & Cook, R. E. (1972). Causal ascriptions and achievement behavior: A conceptual analysis of effort and reanalysis of locus of control. *Journal of Personality and Social Psychology, 21*(2), 148–239.

Wellman, H. M., & Bartsch, K. (1988). Young children's reasoning about beliefs. *Cognition, 30*(3), 239–277.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43*(1), 337–375.

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology, 56*(2), 161–169.

White, P. A. (2014). Singular clues to causality and their use in human causal judgment. *Cognitive Science, 38*(1), 38–75. http://dx.doi.org/10.1111/cogs.12075.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General, 136*(1), 82–111.

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford, England: Oxford University Press.

Yablo, S. (2002). De facto dependence. *The Journal of Philosophy, 99*(3), 130–148.

Young, L., & Phillips, J. (2011). The paradox of moral focus. *Cognition, 119*(2), 166–178. http://dx.doi.org/10.1016/j.cognition.2011.01.004.

Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Counterfactuals and causality in group attributions. *Cognition, 125*(3), 429–440.