Wins above replacement: Responsibility attributions as counterfactual replacements

Tobias Gerstenberg[1] (tger@mit.edu), **Tomer D. Ullman**[1] (tomeru@mit.edu), **Max Kleiman-Weiner**[1] (maxkw@mit.edu)
David A. Lagnado[2] (d.lagnado@ucl.ac.uk) & **Joshua B. Tenenbaum**[1] (jbt@mit.edu)

[1]Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139
[2]Cognitive, Perceptual and Brain Sciences, University College London, London WC1H 0AP

Abstract

In order to be held responsible, a person's action has to have made some sort of difference to the outcome. In this paper, we propose a *counterfactual replacement model* according to which people attribute responsibility by comparing their prior expectation about how an agent was going to act in a given situation, with their posterior expectation after having observed the agent's action. The model predicts blame if the posterior expectation is worse than the prior expectation and credit if it is better. In a novel experiment, we manipulate people's prior expectations by changing the framing of a structurally isomorphic task. As predicted by our *counterfactual replacement model*, people's prior expectations significantly influenced their responsibility attributions. We also show how our model can capture Johnson and Rips's (2013) findings that an agent is attributed less responsibility for bringing about a positive outcome when their action was suboptimal rather than optimal.

Keywords: counterfactuals; responsibility; Bayesian inference; attribution; theory of mind.

Introduction

How do we hold others responsible for their actions? There is a strong intuition that someone can only be held responsible if what they did made a difference to the outcome. There are at least two ways to assess the extent to which a person made a difference to the outcome. The first way is to consider an *action-centered* contrast and compare the actual action a person took with alternative actions she could have taken. Imagine a Dr. Smith who administers a treatment that causes a patient to suffer from severe side effects. If there were alternative options that would have led to a better outcome for the patient, we might blame Dr. Smith for the choice she made. In contrast, if we believe that the alternative options would have led to even worse side-effects, we might think Dr. Smith's decision is creditworthy. The second way to assess difference-making is to consider a *person-centered* contrast and compare what a person did with what other persons would have done in the same situation. We might not blame Dr. Smith for the negative outcome – even if there was an alternative treatment that would have been better for the patient – if we believe that other doctors would have prescribed the same treatment in her place.

In the law, we find both *action-centered* and *person-centered* contrasts. According to the "but-for test", a defendant's action is deemed a factual cause of a negative outcome if the outcome would not have occurred but for his action (Hart & Honoré, 1959/1985). According to the "reasonable man test", whether a person's behavior is deemed negligent depends on what a reasonable man would have done in the given situation (Schaffer, 2010). Another example of a *person-centered* contrast is found in baseball, where a statistic called "Wins Above Replacement" (WAR) expresses the additional number of wins a player contributes to the team's success, compared to the estimated number of wins the team would have achieved with a replacement (Jensen, 2013).

Previous work in psychology has shown that both *action-centered* and *person-centered* counterfactual contrasts influence responsibility attributions. For example, according to Brewer (1977), responsibility attributions are (i) negatively related to the subjective probability that the outcome would have occurred in the absence of the person's action, and (ii) positively related to the subjective probability that the outcome will occur given the person's action (see also Petrocelli, Percy, Sherman, & Tormala, 2011; Spellman, 1997). Fincham and Jaspars (1983) showed in a series of experiments that responsibility attributions are also significantly influenced by the subjective degree of belief that another person would have acted in the same way as the protagonist did in the given situation.

In this paper, we propose a formal model of responsibility attribution that is based on the *person-centered* contrast. Inspired by the "reasonable man test" and WAR, we model people's intuitive judgments of blame and credit in terms of counterfactual replacements. Our *counterfactual replacement model* predicts that people assign blame or credit by comparing their *prior expectations* about how likely a person is to bring about a positive outcome in a given situation, with their *posterior expectations* after they have observed the person's action. The more a person's action changes our expectation for the better, the more credit we give that person for their action (cf. Ajzen & Fishbein, 1975). Conversely, we blame a person to the extent that our expectations decrease after having observed their action.

A key advantage of our model over *action-centered* accounts is the natural way in which it captures how intentional and accidental actions are deserving of different degrees of responsibility (Lagnado & Channon, 2008). Our model predicts that responsibility attributions are a function of the extent to which the observation of a person's action leads us to change our expectations about their future behavior. Intentional actions potentially carry rich information about the person's invariant character traits that are predictive of their future behavior. Accidental actions, in contrast, usually carry less information about the person's character (Heider, 1958).

By explicitly modeling prior expectations about a person's characteristics, our model incorporates normative considerations for action from *action-centered* accounts of responsibility attribution. Most situations in which questions of re-

sponsibility arise carry normative implications for action. For example, previous research has shown that people's responsibility attributions are influenced by moral norms (e.g. Knobe, 2010). Recently, it has also been demonstrated that responsibility judgments are sensitive to norms of rational choice. Johnson and Rips (2013) found that participants attributed more responsibility for a positive outcome to a person when the outcome resulted from an optimal decision, compared to situations in which the same positive outcome resulted from a suboptimal choice.

The rest of the paper is organized as follows: First, we introduce the *counterfactual replacement model*. Then, we demonstrate in a novel experiment how people's responsibility attributions are affected by manipulating the framing of a structurally isomorphic task. This finding cannot be explained in terms of an *action-centered* account but is naturally captured by our model. Finally, we show that our model also provides a compelling account of recent data taken to support an *action-centered* account of responsibility attribution (Johnson & Rips, 2013).

Counterfactual replacement model

We model the assignment of responsibility as a difference of expectations – the difference between the prior expectation E_{prior} that a person's actions are going to lead to positive outcomes in the future, and the posterior expectation $\mathrm{E}_{posterior}$ after observing that person's action. Our model expresses expectations by representing people as agents that take actions according to some decision policy. Before observing a person's actions, the model has a certain belief about the sort of person being observed, and on this basis anticipates likely actions and outcomes.

After observing a person's actions, the model updates its belief about the person it observed, and then uses this informed belief to predict likely actions and outcomes in the future. We predict that a person will be blamed for their action if the model's belief about expected future outcomes in similar situations is *lower* after having observed the person's action than it was before (i.e. $\mathrm{E}_{posterior} < \mathrm{E}_{prior}$). Conversely, we predict that a person will receive credit for their action if the model's belief about future expected outcomes is *higher* than before (i.e. $\mathrm{E}_{posterior} > \mathrm{E}_{prior}$).

We will apply our model to situations in which a person faces a decision under uncertainty. Consider a person who correctly predicted an unexpected winner of a horse race. If we believe that the positive outcome was the result of skillful forecasting we might credit the person. However, if we believe the decision was unreasonable and the person was just lucky, we would attribute little credit, if any.

Our model begins by assuming a space of possible agent *types* $\mathcal{T}$. Each agent type has a decision function, which maps a state in the world s to a probability distribution over possible actions $\mathcal{A}$. An agent's decision function can potentially be quite complex, taking into account her beliefs, intentions, goals, skills, motivation and so on.

For any given 'world' w, the model has a prior belief distribution over the different agent types, $P(\mathcal{T} = t|\mathcal{W} = w)$. A 'world' summarizes our assumptions about about what sorts of personal characteristics are relevant in a given situation, and how likely a person is to have these characteristics. In the worlds we consider, the relevant characteristics are skill and reasonableness. Once the model observes an agent taking an action a, it updates its belief on the type of agent being observed by using standard Bayesian reasoning:

$$P(\mathcal{T}|\mathcal{A},\mathcal{W}) \propto P(\mathcal{A}|\mathcal{T},\mathcal{W}) \cdot P(\mathcal{T}|\mathcal{W}), \quad (1)$$

where $P(\mathcal{A}|\mathcal{T},\mathcal{W})$ is given by the agent's decision function.

Our model predicts judgments of responsibility as the difference between the expected reward $\mathrm{E}[r]$ given a prior distribution over the agent space, and the expected reward given a posterior distribution (informed by the agent's action). So, if we see a person take an action a, our model assigns the following judgment:

$$Responsibility(a) = \mathrm{E}_{posterior}[r|\mathcal{T}] - \mathrm{E}_{prior}[r|\mathcal{T}], \quad (2)$$

where the prior is $p(\mathcal{T}|\mathcal{W})$, and the posterior is $p(\mathcal{T}|\mathcal{A},\mathcal{W})$, calculated according to Equation 1. The expectation is taken over the set of different possible situations in a given world.

While the space of possible agents can in principle be very rich, we restrict ourselves here to a relatively simple space of three possible agent types: *reasonable*, *unreasonable*, and *skilled*. Since the expressiveness of the model grows with the space of agents, we deliberately keep the number of agents small. The space of actual agents that people can consider is no doubt much larger and richer. The decision policies associated with the three agent types are the following:

1. Reasonable: This agent chooses actions probabilistically in proportion to their estimated value, using a 'softmax' weighting function:

$$p(a_i|\mathcal{T} = reasonable) = \frac{exp(\beta \hat{r}_i)}{\sum_j exp(\beta \hat{r}_j)}, \quad (3)$$

where $\hat{r}_i$ is the estimated reward from action a_i, and β is a noise parameter that captures the determinism in an agent's planning. If $\beta >> 1$ the agent will almost always choose the action with the greater estimated reward, that is she will tend to 'maximize'. If β is close to 0 the agent will choose more randomly among her actions ($\beta = 0$ implies choosing at chance). A medium β value roughly corresponds to a 'probability matching' strategy. The 'softmax' function is a standard choice to capture different decision strategies (e.g. Sutton & Barto, 1998; Ullman et al., 2009).

2. Unreasonable: This agent will do the 'opposite' of a reasonable agent.

$$p(a_i|\mathcal{T} = unreasonable) = 1 - p(a_i|\mathcal{T} = reasonable) \quad (4)$$

This agent type is meant to capture agents who for whatever reason are more likely than not to do the wrong thing – they could be confused or silly or foolish, etc.

3. Skilled: The skilled agent is similar to the reasonable agent, except that she estimates her reward using the true state of the world:

$$p(a_i|\mathcal{T} = skilled) = p(a_i|\mathcal{T} = reasonable, \hat{r}_i = r_i^{true}), \quad (5)$$

where r_i^{true} is the actual reward that would result from taking action a_i. This agent can be thought of as being more informed than the reasonable agent. In the previous horse-race example, a skilled agent would bet on the "unlikely" horse, if that horse was in fact going to win a particular race.

Choice of parameters

The parameters in our model are the rewards associated with possible outcomes, the decision noise parameter β, and the prior probabilities on the agent types. We set the difference in reward between a positive and a negative outcomes to be $\Delta r = 1$. For our experiment reported below, we fit β based on participants' own behavior in the task. The only remaining free parameters are the prior probabilities on the agent types.

Experiment

Participants' task in the experiment was to judge to what extent an agent who made a decision under uncertainty 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 (20%, 40%, 60% or 80%), (ii) the outcome (positive or negative) within participants, and (iii) the task framing (penalty kick task or spinner prediction task) between participants.[1]

In both task frames, the agent chose one of two available actions. In the *penalty task*, participants evaluated the actions of goalkeepers in soccer who could either jump to the left or to the right side. The outcome was positive (from the perspective of the goalkeeper) if she decided to jump in the direction in which the striker shot, and negative otherwise. The goalkeeper knew about the striker's tendency to shoot in one direction or the other. For example, a goalkeeper might know that a particular striker tends to shoot to the right 80% of the time and to the left 20% of the time. The goalkeeper also knew that the striker is unaware of the fact that the goalkeeper has this information. Participants were further informed that the goalkeeper had no ability to anticipate the shot.

In the *spinner task*, participants evaluated the actions of contestants in a game show whose task was to predict the outcome of a two-colored spinner. For example, a contestant might know that a spinner's chances of landing on yellow or blue are 80% and 20%, respectively.

We predict that the different task framing will change people's assumptions about the plausibility that an agent could exhibit skill. Consider a situation in which the agent chose the 20% option and succeeded. In the *penalty task*, it is conceivable that the goalkeeper's unlikely success was due to skill. Intuitively, if we believe in the possibility of skill then saving the unexpected ball seems more creditworthy than saving a ball that was shot in the expected direction.

In contrast, in the *spinner task*, it is much less plausible a-priori that the agent's correct prediction of the unexpected outcome is due to skill. It is more likely that the agent acted suboptimally, and so we perceive their action as less creditworthy compared to an agent who acted optimally (cf. Johnson & Rips, 2013).

[1]Demos of all experiments reported in this paper may be accessed here:
http://web.mit.edu/tger/www/demos/prior_demos.html

Methods

Participants and materials 83 participants (39 female, $M_{age} = 33.7$, $SD_{age} = 11.14$) were recruited via Amazon Mechanical Turk.

Design and procedure Table 1 shows the eight different rounds that participants experienced. In each round, participants saw two different situations (contrasted on the same screen) and indicated their responsibility judgments on sliders placed underneath each situation (ranging from 0 to 100). 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. The outcome in both situations was positive, i.e. both goalkeepers saved the ball in the *penalty kick* task or both contestants predicted the correct outcome in the *spinner prediction* task. 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.

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).

After having been introduced to the basic features of the experiment and before evaluating the behavior of either the goalkeepers or game show contestants, participants played the game themselves for ten trials. As goalkeepers, they decided whether to jump in the right or left corner and as game show contestants, they predicted which out of two colors the spinner will land on. On average, it took participants 6.19 ($SD = 1.75$) minutes to complete the experiment.

Results and discussion

Figure 1 shows the results for the penalty task (top) and the spinner task (bottom) separated for judgments of blame (left) and credit (right) together with the predictions of the *counterfactual replacement model*. Because there was no effect of pairing as manipulated between blocks I and II, we aggregated the judgments for the pairs of trials in which the agent's

Table 1: Pairings of decisions and outcomes used in the experiments.

Block	Round	Decision		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%	+

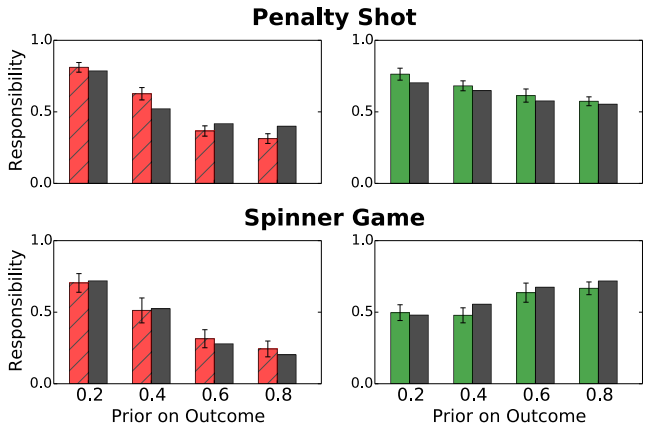

Figure 1: Mean blame (red, striped) and credit (green) judgments for the penalty task (top) and spinner task (bottom) for decisions with different prior chances of being successful. Model predictions are shown in black. Error bars in all figures indicate *SEM*.

decision and outcome were identical (e.g. player 1's decision in rounds 1 and 7).

In both the penalty and spinner task, participants' blame and credit attributions were significantly affected by the extent to which the action was unexpected. The more unexpected an action was, the higher the blame, in both the penalty task ($F(3,120) = 47.92, p < .01$) and the spinner task ($F(3,120) = 27.59, p < .01$).

The relationship between participants' credit attributions and the expectedness of a given action differed between the penalty and spinner task. While in the penalty task, credit ratings *decreased* the more expected an action was ($F(3,120) = 4.45, p < .01$), in the spinner task, credit ratings *increased* for more expected actions ($F(3,123) = 4.06, p < .01$).

We will focus here on how our *counterfactual replacement model* captures qualitatively, the different patterns of attributions between the penalty and spinner task, by assuming that the task framing led to different prior beliefs about the probability that a person could be skilled. While there is a relatively large space of priors over agent types that can account for participants' judgments, the space is different between the penalty and spinner task. For any prior probability placed on the reasonable agent, the best fitting models for the penalty task assign a higher probability on the skilled agent compared to the spinner task. For the model predictions shown in Figure 1, we used a prior over agents of $[reasonable = 0.5, unreasonable = 0.05, skilled = 0.45]$ for the penalty task, and of $[r = 0.5, u = 0.25, s = 0.25]$ for the spinner task.[2]

We fit the β in the agents' decision functions based on participants' own responses in the first part of the experiment, in which they acted as goalkeepers or game show contestants themselves. Participants tended to 'maximize' and predict the more likely outcome most of the time. They predicted the unlikely outcome slightly more so in the penalty task ($mean = 1.02$, β of 7) than in the spinner task ($mean = 0.48$, β = 11). The predictions of the model are not affected by the exact choice of β values, and lead to high correlations with participants' judgments as long as $\beta \gg 1$.

Let us describe intuitively how the *counterfactual replacement model* captures participants' blame and credit judgments. In general, whenever the model observes an agent taking an expected action, it shifts its probability toward the reasonable agent. Conversely, if the agent takes an unexpected action, the model shifts its probability toward the unreasonable agent. The more unexpected the action is, the stronger the predicted shift. Since the skilled agent always correctly predicts the positive outcome, the model assigns zero probability to that agent when a negative outcome occurred. For positive outcomes, the model increases its belief that the agent was skilled. The more unexpected a positive outcome was, the greater the probability that the agent was skilled.

Remember that the expected future reward of the skilled agent is greater than that of the reasonable agent and that of the unreasonable agent. So, our model predicts credit when the posterior belief shifts away from the unreasonable and towards the reasonable and skilled agent. It predicts blame when the posterior shifts toward the unreasonable agent.

For negative outcomes, the model predicts higher blame for unexpected actions than for expected actions. The negative outcome rules out the skilled agent, and unexpected actions are more diagnostic of unreasonable than reasonable agents. This trend – higher blame for unexpected actions – is not reversible for any choice of priors in the model. The model correctly captures that blame attributions increase the more unexpected an action was. The model also correctly predicts that the differences in blame for expected actions (60% vs. 80%) are much smaller.

For positive outcomes, whether the model predicts an increase or decrease in credit for unexpected compared to expected actions depends on the prior over the agents. In the penalty task, the model puts a much higher prior on the skilled than on the unreasonable agent. So, an unexpected successful action increases the model's belief in the skilled agent more strongly than in the unreasonable agent. Upon observing an expected action, the model increases its belief in both the reasonable and skilled agent. However, compared to the unexpected positive actions, the belief in the skilled agent does not increase as much and so the model predicts a decrease in credit the more expected a person's action was.

In the spinner task, the model puts a lower prior on the skilled agent than in the penalty task. So, unexpected actions increase the model's belief in the skilled and unreasonable agent to a similar degree. Expected actions, in contrast, increase the belief in the reasonable agent. Compared to unexpected actions, expected actions lower the belief in the unreasonable agent much more strongly than the skilled agent. Given this, the model overall predicts that credit *increases* for expected positive outcomes compared to unexpected ones.

[2] We used the following linear transformation to fit the model's responses m to the scale used by participants: $a + b \times m$. The parameters were $[a = -0.22, b = 1.5]$ for blame and $[a = 0.5, b = 1.85]$ for credit.

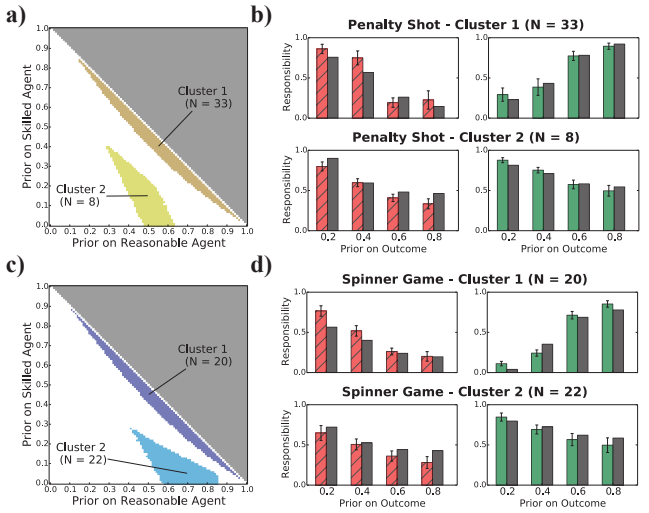

Figure 2: Left panel: Region in the prior probability space over agents for which the model yields a $RMSE < 10\%$ for the two different participant clusters. Gray regions represent impossible priors. Right panel: Mean blame (red, striped) and credit (green) judgments together with the model predictions (black) separated for two clusters of participants in the penalty (top panel) and spinner task (bottom panel).

To sum up, whether the model predicts a positive or negative relationship between credit and the extent to which an agent's action was expected depends on the prior we assign over the different agents, and on how the posterior over agents shifts from the prior based on the inferences we can make about what agent is most likely to have generated the positive outcome.

Individual differences Our model suggests that the difference between responses in the penalty and spinner cases is driven by how much participants believe a-priori that the agent they are judging is likely to be skilled. But is this belief uniform, or are there sub-groups within conditions, and if so how many? To test this, we applied a Gaussian Mixture Model with a varying number of possible Gaussians to cluster participants' responses for each the penalty and spinner tasks. The Cross-Validated Likelihood method (Smyth, 2000) showed that the data for both conditions are best clustered into two groups each.[3]

The average response by cluster for the different cases is shown in Figure 2b and d, together with the predictions of our model for each cluster. Interestingly, in both cases the two groups of participants correspond to the following: there are those who give more credit the more expected an action was to succeed, and those who give less. It is the relative proportion of these groups that varies between the two conditions. Again there is a large space of possible priors that can fit these results, and it is not the exact value but the relation between the priors that is interesting. In Figure 2a and c we show the sub-space of priors that best captures each cluster. The colored patches correspond to the regions of priors that yield the best fit to the different clusters (the 10% best RMSE scores). The region that best fits Cluster 1 in the penalty case corresponds to a greater belief in skill than the region that best fits Cluster 2. The same is true for the spinner case. For the model predictions shown in Figure 2b and d, we simply used the center of the best-fitting prior regions as shown in Figure 2a and c.

So, our model suggests that the two clusters of participants' responses (found independently of the model), are best explained by a different a-priori belief in the skill of the agent. In both the penalty and spinner task, participants are split into two groups: One group believes that the agent is probably unskilled, and therefore less deserving of credit for choosing the unexpected action that happened to succeed. The other group believes the agent is potentially skilled, and therefore worthy of credit when acting unexpectedly and succeeding. In the penalty case the ratio is about 4:1 in favor of those who place higher probability on skill, while in the spinner case the ratio is about 1:1. The difference in the cluster sizes between the two conditions is reflected in the best-fitting prior for the aggregate results shown in Figure 1.

Modeling Johnson and Rips's (2013) results

Our model accounts well for people's individual and aggregate responses, but can the same pattern be predicted by existing models? Johnson and Rips (2013) proposed an *optimality model* according to which responsibility attributions are higher for optimal than for suboptimal choices. The model further predicts that there are no differences within optimal and suboptimal choices. The optimality model fails to explain two key effects in our data. First, it does not account for the increase in blame the more suboptimal a person's choice was. Second, it cannot predict the shift of the relationship between attributed credit and an action's chance of leading to a positive outcome between the penalty and spinner task. In the following, we will show that our *counterfactual replacement model* provides an alternative account of their findings.

Johnson and Rips (2013) had participants judge the extent to which an agent was responsible for bringing about an outcome E. In all cases, the agent knew the chances with which two different actions would lead to a desired outcome. The agent always chose option A which had a 50% chance of causing the positive outcome. Between situations, Johnson and Rips (2013) varied the probability with which the alternative option B would have brought about the outcome. They found that participants attributed more responsibility when the agent had made the optimal choice (i.e. when $p(E|A) > p(E|B)$) compared to situations in which her choice had been suboptimal (i.e. $p(E|A) < p(E|B)$).

Model predictions As we do not have participant data to empirically fit β, we set $\beta = 9$, in between the empirical values for the penalty and spinner cases. However, the analysis is insensitive to the exact choice of β as long as $\beta \gg 1$.

Figure 3 shows participants' responses as well as our

[3]With 1000 cross-validation runs, the mean log-likelihood score for using 1–4 clusters, relative to the score for two clusters, was $(-1.79, 0, -4.94, -11.94)$ for the penalty task and $(-16.39, 0, -2.93, -8.01)$ for the spinner task. The two cluster solution proved to be stable with participants being grouped into the same clusters 99% of the time.

model predictions, re-scaled to match the scale used by Johnson and Rips (2013). In line with previous experiments, we used a prior of 0.5 for the reasonable agent and fit the proportion of skilled and unreasonable agent resulting in the following prior: $[r = 0.5, u = 0.5, s = 0]$. This analysis suggests that participants didn't believe that skill was a likely possibility in the Johnson and Rips (2013) task.

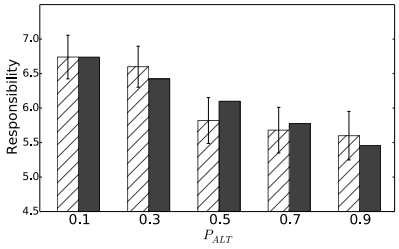

Figure 3: Mean responsibility judgments (white bars) for Experiment 1A in Johnson and Rips (2013), and our model predictions (gray). P_{ALT} is the probability that the alternative choice would have succeeded.

While Johnson and Rips's (2013) *optimality model* predicts a 'step-function' which does not differentiate between the degrees to which a decision was optimal or suboptimal, our model does predict a linear trend. The *optimality model* does not explicitly predict responsibility values for situations in which both choices are equally optimal. Our model predicts that in such cases the responsibility should be between the values assigned for optimal and suboptimal choices.

General discussion

In this paper, we proposed a *counterfactual replacement model* of responsibility attribution. According to this model, attributions of blame or credit are a function of how much observing a person's action changes our belief about the likely future actions and rewards of that person. The model predicts credit to the extent that our expectations about the person's behavior are increased after having observed a person's action, and blame to the extent that they decrease.

Our model draws from a rich tradition of ideas in attribution research such as Bayesian belief updating (Ajzen & Fishbein, 1975), the diagnostic value of different actions for dispositional inference (Heider, 1958), and the extent to which observed behavior differs from expectations (Fincham & Jaspars, 1983). By defining people's prior beliefs as a concrete hypothesis space over possible agents, our model makes these ideas precise and yields testable quantitative predictions.

In a novel experiment, we showed how manipulating people's prior beliefs about the plausibility of skill influenced responsibility attributions. Participants who believed that skill was a relevant factor, attributed more credit for unexpected positive outcomes than for expected ones. In contrast, participants who doubted the plausibility of skill, saw unexpected outcomes as diagnostic for unreasonable behavior (rather than skill) and so attributed less credit than for expected positive outcomes. We further found that there were systematic inter-individual differences between participants within our experimental conditions and demonstrated how these differences can be captured by our model in terms of differences in prior beliefs. Finally, we also showed how our *counterfactual replacement model* accounts for the finding that people attribute less responsibility for bringing about a positive outcome via a suboptimal rather than an optimal choice (Johnson & Rips, 2013).

Our model extends Johnson and Rips's (2013) optimality model and clarifies the way in which action expectations and considerations of optimality influence responsibility attributions. Rather than predicting a direct mapping from the expectedness of a person's action to the responsibility judgment, we predict that this mapping is mediated by a situation-dependent inference about the person's character. Whether an unexpected action leads to more or less responsibility compared to an expected action depends on whether the action is diagnostic for skill or unreasonableness in the given context. In future research, we will continue to explore how considerations about what actually happened and expectations about future behavior interact to determine blame and credit attributions. In terms of our framework, the intercept of the linear transformation described in Footnote 2 can be interpreted as reward (or punishment) for the particular situation while the slope determines to what extent blame and credit are influenced by updated expectations about future behavior.

Acknowledgments

TG, TDU, and JBT were supported by the Center for Minds, Brains and Machines (CBMM), funded by NSF STC award CCF-1231216 and by an ONR grant N00014-13-1-0333. MKW was supported by a Hertz Foundation Fellowship and NSF-GRFP.

References

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

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

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

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

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

Jensen, S. (2013). A statistician reads the sports pages: Salaries and wins in baseball. *CHANCE*, *26*(1), 47–52.

Johnson, S. G. B., & Rips, L. J. (2013). Good decision, good causes: Optimality as a constraint on attribution of causal responsibility. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2662–2667).

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*(4), 315–365.

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

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.

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

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, *10*(1), 63–72.

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

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

Ullman, T. D., Tenenbaum, J. B., Baker, C. L., Macindoe, O., Evans, O. R., & Goodman, N. D. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in Neural Information Processing Systems* (Vol. 22).