Go fishing! Responsibility judgments when cooperation breaks down

Kelsey Allen (krallen@mit.edu), **Julian Jara-Ettinger** (jjara@mit.edu), **Tobias Gerstenberg** (tger@mit.edu),
Max Kleiman-Weiner (maxkw@mit.edu) & **Joshua B. Tenenbaum** (jbt@mit.edu)
Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Abstract

Many social judgments hinge on assigning responsibility to individuals for their role in a group's success or failure. Often the group's success depends on every team member acting in a rational way. When someone does not conform to what others expect of them, cooperation breaks down. We present a computational model of responsibility judgments for individuals in a cooperative setting. We test the model in two behavioral experiments where participants were asked to evaluate agents acting in a cooperative, one-shot game. In Experiment 1, we show that participants' action predictions are consistent with a recursive reasoning model. In Experiment 2, we show that people's assignments of blame are influenced by both an agent's presumed rationality, or adherence to an expected policy, as well as the pivotality of the agent's actions, or how close the situation was to one in which the action would have made a difference to the outcome.
Keywords: responsibility attribution; theory of mind; recursive reasoning; multi-agent coordination.

Introduction

Imagine that you are a fisherman living in a remote village in the Amazonian rainforest. Your village survives by trading fish with neighboring groups who visit each day, and then distributing the profit amongst all villagers. One morning, you wake up to find out that the only road into your village is blocked by three trees that fell during an overnight storm. Someone needs to clear the road or else your village will be unable to trade today. You know most of the fishermen are stronger than you, and certainly strong enough to move the trees without your help before traders arrive. Since it is in everyone's best interest to clear the road, you assume that the stronger fishermen will clear the road, and you head out early to fish. When you come back with the day's catch, you discover that the road is still blocked. Everyone went fishing and assumed that someone else would clear the trees. Who's to blame?

Assigning responsibility when a team's efforts go right or wrong is an essential element of social life. Our goal in this paper is to propose and test a new computational model for these responsibility judgments in a cooperative setting. Previous psychological accounts of credit and blame (Lagnado, Gerstenberg, & Zultan, 2013; Gerstenberg, Ullman, Kleiman-Weiner, Lagnado, & Tenenbaum, 2014; Spellman, 1997) have identified two broad factors as important in evaluating agents and their actions. The first are **person-centric** (Gerstenberg et al., 2014), based on expectations about how people are likely to act, or norms of how they should act in a given setting. Someone is blamed more to the extent that they failed to act the way they were expected to. This motivates a consideration of *rationality* as capturing an agent's ability to plan according to an appropriate norm (Johnson & Rips, 2015). The second are **action-centric** judgments (Spellman, 1997), based on retrospective evaluations of how much an action contributed to a good or bad outcome. A specific action receives more blame to the extent that it made a negative difference to the team's outcome.

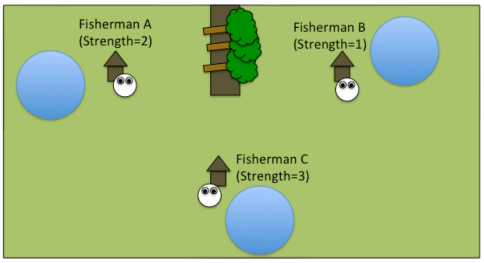


Figure 1: Set-up of three fishermen in a fishing village with a road blocked by three trees.

Our work is in part inspired by Lagnado et al. (2013), who proposed a specific model for these two factors in the context of team actions with all-or-nothing reward, i.e., the team either succeeds or fails. They captured action-centric responsibility with a counterfactual measure they called "pivotality", and person-centric responsibility with a measure they called "criticality". We find that in extending this approach to cooperative action with graded potential rewards, where the team can succeed to a greater or lesser extent, both of these notions have to be generalized. Pivotality is relatively straightforward; in the example above, only the strong fishermen were pivotal, because only if they had chosen to clear the road would the outcome have been different. The most interesting new contribution of our work is in assessing the person-centric aspect of responsibility. We find that rationality, or the assumption that your teammates will do what you expect them to do, influences people's responsibility judgments.

Intuitively, rationality is a key component of blame attribution for many everyday cooperative tasks. If you are distributing bonuses to employees at an investment firm, you may not want to give as much money to a broker whose strange decisions cost the company revenue. A coach who made a bad call instructing the quarterback to pass the football instead of running it up the field might be blamed more for the team's failure than the receiver who didn't catch the ball. To illustrate the importance of rationality in our fishermen example, imagine once again you see three trees blocking the road. In this life, you are strong, so you could either go clear all three trees, or collect three fish sacks. Your two friends Arnold and Bob, however, are weaker. Bob can either clear one tree from the road, or collect one fish sack, while Arnold can clear two trees, or collect two fish sacks (see Figure 1). Because you know that neither of your friends is strong enough to clear the

road themselves, you choose to go clear the trees and expect your friends will go fishing. However, when you get to the road you find that Arnold is also there, and it's too late now for him to go fishing. The road is cleared at the end of the day, but your village ended up trading only one fish sack (that Bob collected). Arnold's choice didn't cause the group to fail, but it nevertheless turned out to have been a bad choice. What matters here is not just the pivotality of each action in hindsight, but also each agent's rationality at the stage when the actions were planned. If Arnold had reasoned similarly to you, he would have realized that you would clear the trees, and therefore he would have gone fishing. It was therefore his inability to predict the actions of the other agents in the group and plan accordingly which led to the group receiving less than the ideal outcome.

The remainder of the paper is organized as follows. We first develop and experimentally verify a model of how agents in this coordinative, one-shot game should act under various configurations of fishermens' strengths and number of trees. We then show that both person-centric *rationality* and action-centric *pivotality* are important aspects of blame attribution when the fishermen are not able to achieve their optimal outcome. Finally we suggest follow-up experiments to test the sensitivity of human judges to optimality, as well as investigations of credit attribution and how judgments change over time when there are repeated interactions between fishermen.

Computational Models

We use the experimental paradigm outlined in the introduction and consider three fishermen (A, B and C) living in the village. They live far away from each other, each near a pond in which they can fish. There is also a road entering the village which is blocked by either one, two or three fallen trees (referred to as $T = 1,2,3$). The fishermen each have an associated strength (between one and three, referred to as $S(A)$, $S(B)$ and $S(C)$) which corresponds to how many sacks of fish they can obtain from one day of fishing, or the number of trees they can clear from the road. The scenario from Figure 1 would therefore be represented as $T = 3, S(A) = 2, S(B) = 1, S(C) = 3$. At the end of the day, if the road has been cleared, the fishermen equally distribute the money earned from the fish sacks they have collected. If the road is not cleared, they receive nothing.

We first develop two possible models of rational action selection for a fisherman in this paradigm. After discussing the models of rational action, we consider two models of pivotality, and suggest that blame judgments are related to violations of expectations as well as pivotality considerations.

Model of action

In a purely cooperative coordination game, individuals should attempt to find an optimal strategy to maximize the expected reward of the group (Schelling, 1980). If there is only one way for the group to succeed, and you know all group members are rational, you can choose your action without worrying about what the others will do. However, when there is more than one way for the group to get the optimal reward, and these have conflicting strategies for each agent, the choice is less clear.

Recursive reasoning with soft-max We model the uncertainty in this decision making process by considering rational agents who each try to best respond to their companions at a level k depth of reasoning (Yoshida, Dolan, & Friston, 2008). We can then define the probability of a fisherman i taking action a_i at a depth of reasoning k according to a soft-max on his expected reward for action a_i. This involves two steps: first calculating the probabilities for the actions of the other agents at a level $k-1$, and then choosing a response that maximizes your own expected reward under these probabilities:

$$p^k(a_i) = \frac{\exp(\beta \hat{r}_k[a_i])}{\sum_{a_i \in actions} \exp(\beta \hat{r}_k[a_i])} \quad (1)$$

$$\hat{r}_k[a_i] = E_{-i_{k-1}}[R|a_i] \quad (2)$$

$p^k(a_i)$ is the probability, at level k, that fisherman i should take action a_i. R is the reward table describing the number of fish sacks sold by the fishermen under each combination of actions. $R|a_i$ is then the subset of rewards where fisherman i took action a_i. $\hat{r}_k[a_i]$ is the expected reward at level k of action a_i, calculated using $p^{k-1}[a_{-i}]$. Finally, β is a rationality parameter describing how likely the agent is to choose a random action (with $\beta = 0$ being completely random, and $\beta >> 1$ corresponding to always choosing the action which gives the maximal expected reward).

Alternative uniform choice over optimal strategies A reasonable alternative model might be to consider agents who choose an action uniformly from those which might lead to an optimal reward. In the case $T = 3$, $S(A) = S(B) = 1$, and $S(C) = 2$, this model would predict a 50% likelihood for fisherman A to clear the trees and a 100% likelihood for fisherman C to clear the trees. Here, there are two sets of actions leading to optimal reward: fisherman A fishes while B and C clear the road, or fisherman A and C clear the road while B fishes. In both situations, fisherman C must clear the road, and so his action is clear. However, for fisherman A, there is one scenario in which he should fish, and one in which he should clear the trees, so this model predicts that he will choose either action with 50% likelihood.

Model of blame

Now that we have a model for how agents *should* choose an optimal action in a given scenario, we define the "rationality" aspect of blame as an expectation violation. Mathematically, this is 1 - $p(a_i)$, one minus the rational-action probability of the action a_i that the agent took (fishing, or clearing the road). When it was perfectly clear what action an agent should have chosen ($p(a_{fish}) = 1$ for example), then the agent should receive full blame if he cleared the road, and 0 blame if he went fishing. However, this model completely lacks any consideration of the other agents' actions. In hindsight, perhaps one

of the fishermen made the wrong choice but it didn't matter, because another fisherman also made a bad choice. However, if the other fishermen made the right choices, and only one did not (and he cost the group a lot!) then he may be seen as more to blame. For example, consider the case of $T = 2$, $S(A) = S(B) = 1$ and $S(C) = 3$. Imagine first that fisherman C goes fishing, and fisherman B goes to clear the trees. We may blame fisherman A more for fishing than we would have if fisherman B had also gone fishing. This is captured by the pivotality measure discussed briefly in the introduction. The pivotality of a person's action for a specific outcome in a situation is defined as:

$$Pivotality = \frac{1}{N+1} \tag{3}$$

where N is the minimum number of other agents whose actions need to be changed to make the reward outcome counterfactually dependent on the fisherman in question. In cases where the fisherman made the right choice, but his colleagues failed to do so, pivotality would be 0. A fisherman's pivotality would be 1 if he needed to act differently for the group to receive a reward.

In our scenario, there are discrete rewards, rather than merely binary as in Lagnado et al. (2013). We therefore looked at two modifications to this structural pivotality measure: a distance to the closest optimal strategy, or a distance to the closest strategy where any reward was received.

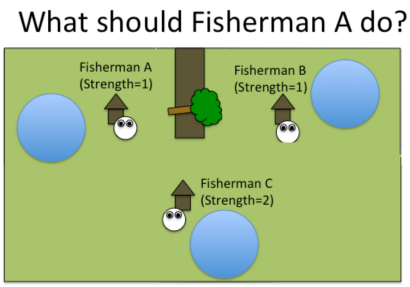
Distance to optimal Pivotality is measured as the distance to the closest optimal strategy.

$$Pivotality_{optimal} = \frac{1}{N_{optimal}+1} \tag{4}$$

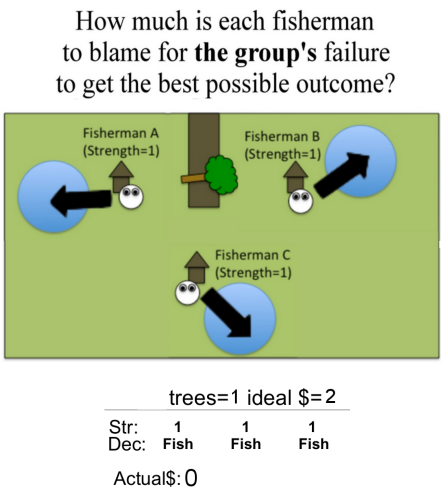
Consider the case where $T = 3$, $S(A) = 2$, $S(B) = 1$, and $S(C) = 3$. This configuration has two strategies leading to maximum reward: either both fisherman A and fisherman B clear the trees while fisherman C fishes, or fisherman C clears the trees while fishermen A and B both fish. Now consider the scenario when only fisherman A went to clear the trees, while both fishermen B and C fished. In this case, the closest optimal strategy is the one in which fisherman B changes his action to clear the trees. Therefore, the pivotality for fisherman A is 0 (in the closest optimal world, he should have done what he did), while the pivotality for fisherman B is 1, and for fisherman C is 0 (like A, his action in the closest optimal world is the same as his actual action). If fisherman C had also chosen to clear the trees, then the new closest optimal strategy would be when fisherman A's action is switched, leading to pivotality scores of 1 for fisherman A, 0 for fisherman B, and 0 for fisherman C.

Distance to any reward In this version of pivotality, instead of considering the closest optimal strategy, we consider any strategy in which the agents would have received some reward.

$$Pivotality_{any} = \frac{1}{N_{any}+1} \tag{5}$$



(a) Experiment 1. Participants were asked to judge fisherman A's best action.



(b) Experiment 2. Example image for blame attribution. Underneath the image is the textual representation of this scenario.

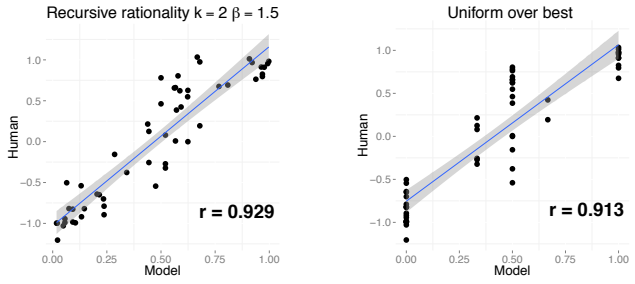
Figure 2: Example images from the two experiments.

Consider the scenarios laid out above for optimal pivotality (for $T = 3$, $S(A) = 2$, $S(B) = 1$, and $S(C) = 3$). In the first case where fisherman A clears the trees, fisherman A would still have pivotality 0, but both fishermen B and C would have pivotality 1 (because either of them could have acted to obtain reward). In the second case where fishermen A and B clear the trees, everyone would have pivotality 0 because they received a nonzero reward (and therefore their policy was satisfactory). This model effectively downweights the blame for agents in any situation where they received reward, and heavily penalizes stronger agents in cases where a weaker agent (or combination of weaker agents) should have gone to clear the trees (like the $T = 2$, $S(A) = 1$, $S(B) = 1$, $S(C) = 3$ case, where fisherman C could have cleared the trees to obtain a suboptimal reward).

We will discuss four models of blame attribution which differ in terms of what aspects they consider: rationality alone, optimal reward pivotality alone, any reward pivotality alone, and a linear mixture of rationality and optimal pivotality given by a weight w.

Experiments

In the first experiment, we asked participants to judge which action fisherman A should take on a sliding scale from "Definitely fish" to "Definitely clear road" (see Figure 2a). They



(a) The soft-max recursive reasoning model ($k = 2, \beta = 1.5$).

(b) Uniform action selection from optimal strategies.

Figure 3: The two models of action selection for Experiment 1.

were given a tutorial explaining the fishermen's situation (similar to the introduction of this paper), and asked to answer some comprehension checking questions. We generated different situations by considering all unique permutations of 1-3 trees and three fishermen with strengths 1-3, leading to 54 different scenarios. Participants were then shown a randomly selected subset of 27 of these. 50 participants were recruited through Amazon Mechanical Turk, giving 25 judgments for each trial.

In the second experiment, we asked participants to judge how much each fisherman was to blame for the group's failure to get the best possible outcome (see Figure 2b). The actions of the fishermen were represented as arrows either towards their pond, or towards the trees on the road. Participants were additionally shown the number of fish sacks which the fishermen actually collected, as well as the best possible number they could have collected, next to the image. The blame for each fisherman was assessed on a sliding scale from "Not at all" to "Very much". Participants were additionally required to go through an introductory tutorial, answer comprehension testing questions, and give optimal strategies for 7 example scenarios (of which they needed to answer 6 correctly to continue).

Since there are many possible combinations of strengths, trees, and choices, we selected only a subset of trials falling into 4 distinct categories. The first category consisted of those trials where all agents chose to go fishing. These were chosen by ordering trials according to participants' average judgments from Experiment 1, and then selecting every fifth element of the resulting ordered list, leading to 10 such trials (Figures 6a, 6d, 6e and 6j). The second category consists of 12 scenarios in which at least one fisherman went to clear the trees, but the fishermen failed to collect any reward, and this was due to their collective failure to clear the fallen trees (see Figures 6f and 6h). For comparison, we also included 8 cases where no fishermen cleared the trees.

The third category includes 15 cases in which the amount of reward received was non-zero, but sub-optimal, and it was not clearly one agent's fault (because there were multiple best responses, like in Figures 6b and 6c). Finally, the fourth category also consisted of 18 cases with sub-optimal reward outcomes, where the action of one agent in the group was more

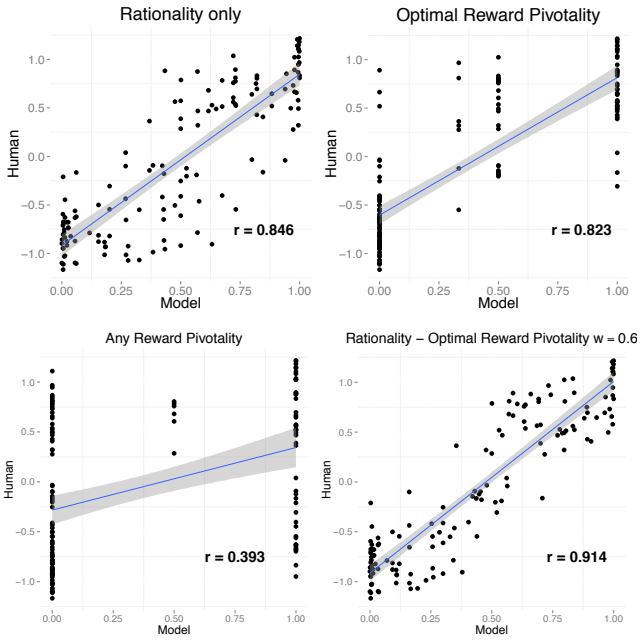


Figure 4: Four models of blame attribution across all 140 fishermen scenarios

clearly incorrect. This category also included intriguing cases such as those where everyone made the incorrect choice, but some reward was still received (like in Figure 6k).

During the experiment, participants were shown 21 out of the 63 total trials. We recruited 60 participants through Amazon Mechanical Turk to participate in this study, leading to 20 judgments per trial.

Results

We first used the data collected in Experiment 1 to determine which model of action selection best predicts participants' judgments. In order to account for individual subjects interpreting the slider's values differently, we z-scored within subjects before averaging and comparing to the model predictions.

As seen in Figure 3, both models fit the participant data well with respect to the correlation coefficient. However, the uniform action selection over optimal policies model has some large outliers. These outliers correspond to situations such as T=3, S(A)=3, S(B)=1 and S(C)=2, where there are two optimal policies, but one of these requires less coordination by the fishermen to clear the trees. In this case, the uniform optimal policy model would say that fisherman A should clear the trees only 50% of the time. However, the recursive reasoning model suggests that he should clear the trees 93% of the time under the fitted parameters. Participants state that fisherman A should clear the trees 82% of the time. The difference between the predictions of these models results from the importance of reasoning about other agents when cooperation is key. Fitting the recursive rationality model to the z-scored participant data using a least-squares regression yields a value for k of 2 and β of 1.5.

As the main contribution of this work, we assess the impor-

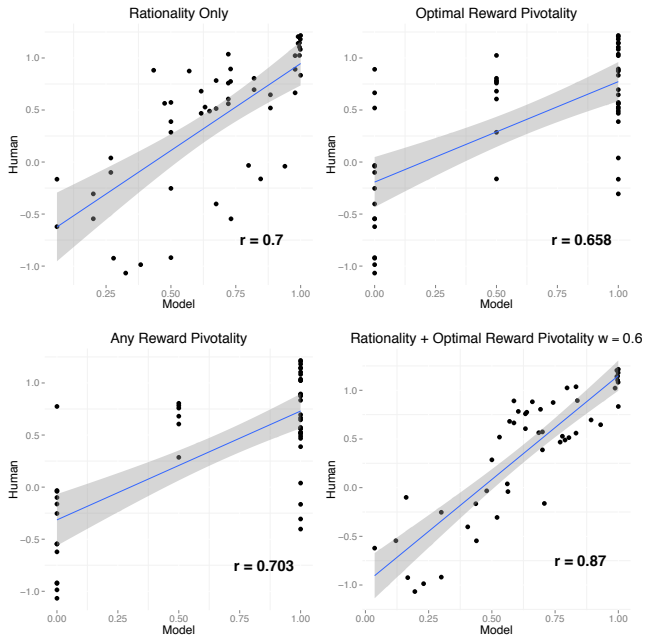


Figure 5: Four models of blame attribution across those cases where no reward was received.

tance of rationality and pivotality for blame attribution when the fishermen do not collect the optimal reward. There are 63 separate scenarios where all fishermen are judged by a participant for a given trial. In total this yields 140 unique judgments of fishermen, for which we have 20 data points each. As in Experiment 1, we z-scored the data on the level of individual participants before averaging their judgments. All model fits were done using a coarse-grained search for k, β and w where appropriate minimizing the residuals from a linear regression between the z-scored human data and the model predictions.

Scatter plots of model predictions and participants' average judgments for four versions of the model are shown in Figure 4. The rationality model has two fitted parameters: k and β. With $k = 2$ and $\beta = 1.9$, the best-fitting parameters for this model are similar to the values found for Experiment 1, and consistent with the rationality + optimal pivotality model as well. Neither the optimal pivotality model nor the any reward pivotality model have any fitted parameters, and fit the data significantly worse than the mixture model. The mixture model has an additional fitted parameter w which corresponds to a linear weighting between rationality and pivotality ($blame = w \times rationality + (1-w) \times pivotality$). The best fit is w=0.60 using the optimal pivotality measure, suggesting an almost equal contribution of rationality and pivotality for blame attribution. Replacing the "optimal pivotality" with the "any pivotality" yields a worse fit.

In order to determine more precisely what the pivotality and rationality models individually capture, we looked at several representative cases in Figure 6, comparing human judgments to the rationality only, and rationality mixture model (which were the only models to give graded responses across the scenarios). The mixture model better accounts for scenarios in which at least one fishermen went to clear the road, such as those shown in Figure 6f and Figure 6h (see Figure 5 for fit). Additionally, in highly unusual cases where all the fishermen made bad decisions (such as that shown in Figure 6j), participants are clearly sensitive to the optimal reward outcome rather than a suboptimal but fairly good reward. However, both of the models overpredict how much blame fisherman C will receive in Figure 6i. This is likely due to participant's sensitivity to fisherman C being responsible for any reward being received, which is common across other similar cases. In this instance, although the pivotality for fisherman C is 0, the rationality model predicts that fisherman C should have gone fishing, because he could have reasonably assumed one of his companions would have cleared the road. Therefore, the "right" decision for receiving reward was actually less rational. For many of these scenarios, the "any reward" pivotality measure is a much better indicator of human blame judgments, although when considering all cases, it still performs significantly worse than the optimal reward pivotality measure.

Examining the cases where the fishermen received nothing due to their inability to coordinate clearing the road yields further insights into the importance of pivotality (Figure 5). Under this set of examples, the "any reward" pivotality model's correlation jumps from 0.39 (when we considered all trials) up to 0.70 across only these cases. This relatively high correlation is driven by the endpoints (where fishermen received either full or no blame). However, combined with the analysis of individual scenarios, it seems that participants are more sensitive to decisions which would change the reward outcome to 0 or from 0 rather than some suboptimal but nonzero outcome. In these trials, the difference between the "rationality only" model and the mixture model also becomes statistically significant, demonstrating the heightened importance of pivotality for these cases.

Discussion

Overall, participants find both person-centered aspects (in the form of rationality based on an expected action), as well as action-centered aspects (optimal pivotality) to be important when assessing the blame of agents in a coordinative game. Unlike previous experiments in responsibility attribution, this paradigm critically incorporates an agent's ability to plan an appropriate action as important for assigning blame. Because the fishermen aren't able to communicate with each other, their planning has to rely on their intuitive theory of how others are going to act in the given situation. Our results suggest that people assume the norm is for each fisherman to reason in the same way - namely as a recursive model in which each fisherman tries to model what actions the others will take.

These observations suggest several different directions for future work. First, we will look at credit attribution when the fishermen are able to split their work between tree clearing and fishing, keeping half of the fish they catch for themselves

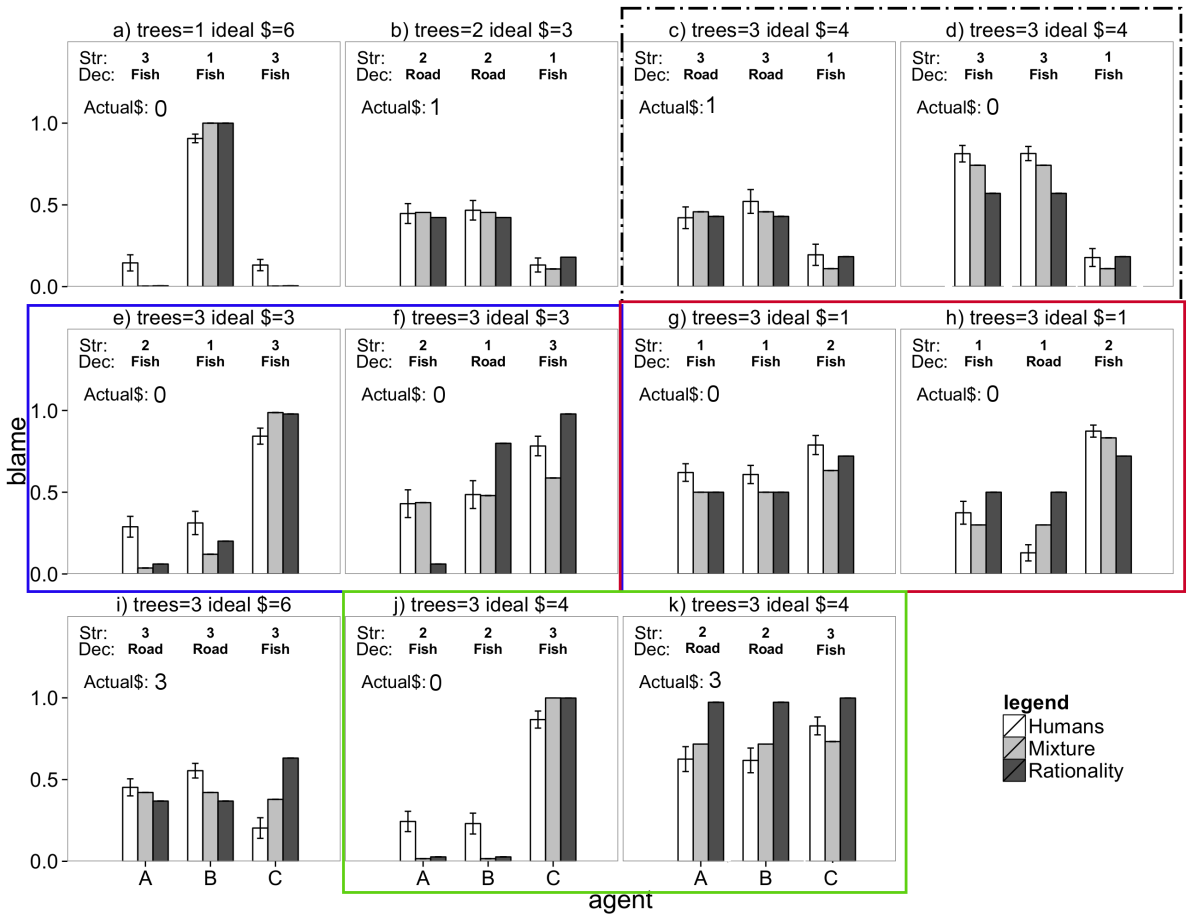


Figure 6: Mean blame judgments (white bars) and model predictions (gray bars) for a selection of different trials. Error bars indicate ± 1 *SEM*. Note: Str = Strength of each fisherman; Dec = Decision to go fishing or clear the trees; ideal = ideal reward; actual = actual reward.

(for example). Second, we will investigate settings in which some of the agents may have negative intentions, or responsibility attribution from the perspective of an agent with different goals from the group (like feeding a very large family).We will incorporate the insights gained from these experiments with work on inverse planning for determining agent's goals and intentions (Baker, Saxe, & Tenenbaum, 2009; Ullman et al., 2010), to capture the "person-centric" aspect of responsibility attribution.

In future experiments, we will also look at a wider range of strengths and trees. Consider the case of T = 1, S(A) = 90, S(B) = S(C) = 1. Here, the difference between suboptimal and optimal reward is more extreme than any of the cases we presented and therefore we may expect a larger range of responses.

Finally, we will extend the current scenario to consider repeated interactions between the same fishermen. Repeated interactions help to establish norms that can guide future action selection (like where the fishermen have settled on a solution with one of two similarly strong fishermen being the tree-cutter).

Acknowledgements

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. MKW was supported by a Hertz Foundation Fellowship and NSF-GRFP.

References

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

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

Gerstenberg, T., Ullman, T. D., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2014). Wins above replacement: Responsibility attributions as counterfactual replacements. In *Proceedings of the 36th annual conference of the cognitive science society.*

Johnson, S. G. B., & Rips, L. J. (2015). Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive Psychology*, *77*, 42–76.

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

Schelling, T. C. (1980). *The strategy of conflict*. Harvard university press.

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

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., & Tenenbaum, J. B. (2010). Help or Hinder: Bayesian Models of Social Goal Inference. *NIPS*, *22*, 1874–1882.

Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. *PLoS Computational Biology*, *4*(12).