Go fishing! Responsibility judgments when cooperation breaks down

Kelsey Allen, Julian Jara-Ettinger, Tobias Gerstenberg, Max Kleiman-Weiner, Joshua B. Tenenbaum

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA

Introduction

- How do we assign responsibility to individuals in a group?
- This question is particularly important when we decide to embark on future research collaborations, give bonuses to employees, and choose a soccer MVP.
- Here we present a computational model of blame attribution in a cooperative one-shot game and test it in two behavioral experiments with adults.

Two aspects of responsibility

Rationality – Person centric

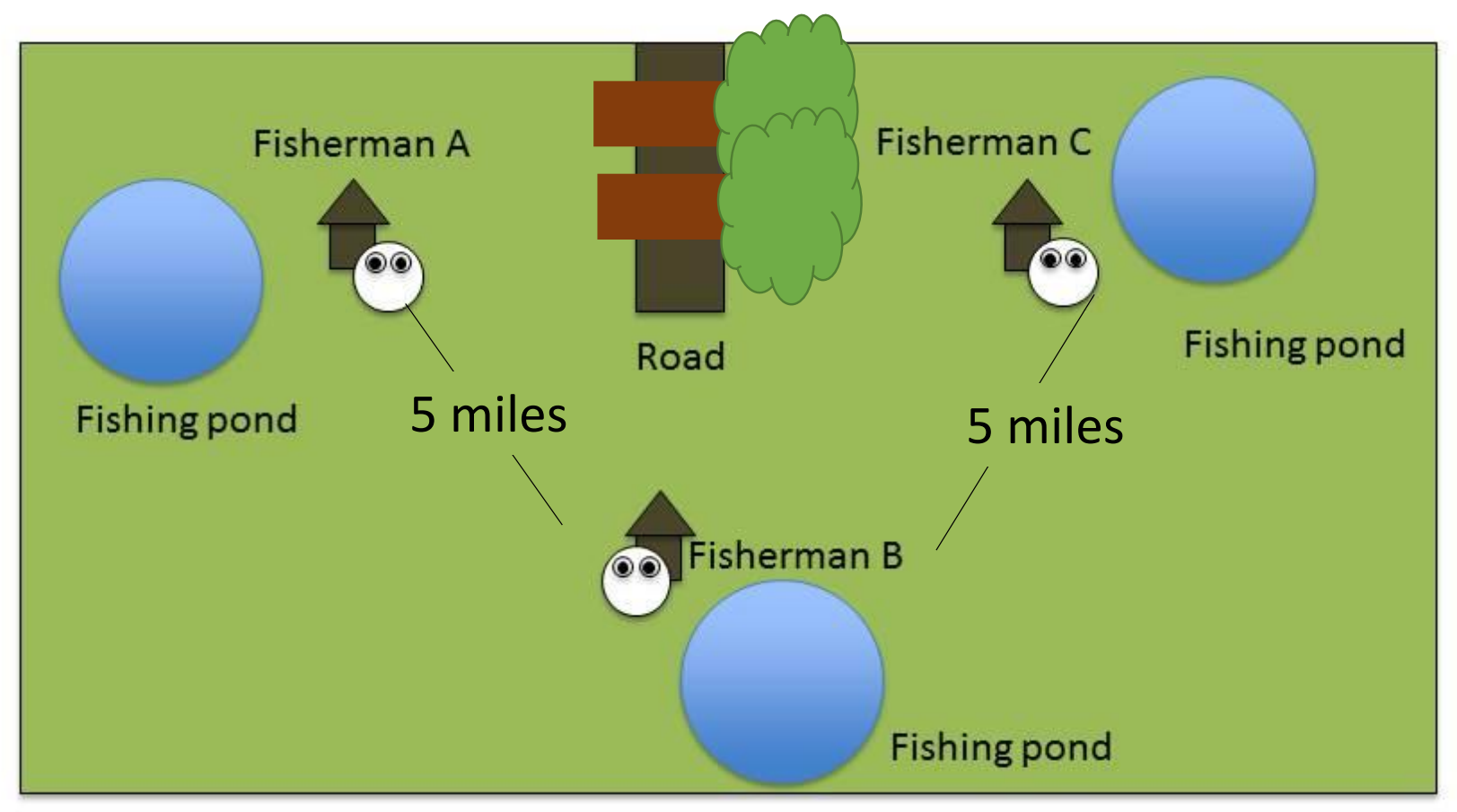
- Agents with good **foresight** should be able to predict the correct action given their knowledge about the world.
  - $Blame = 1 - p(action^*)$
- A measure related to the **agent** and their **reasoning ability.**
- The optimal action (*action**) for an agent will depend on their situation, as well as their individual capabilities with respect to the group.

Pivotality – Action centric

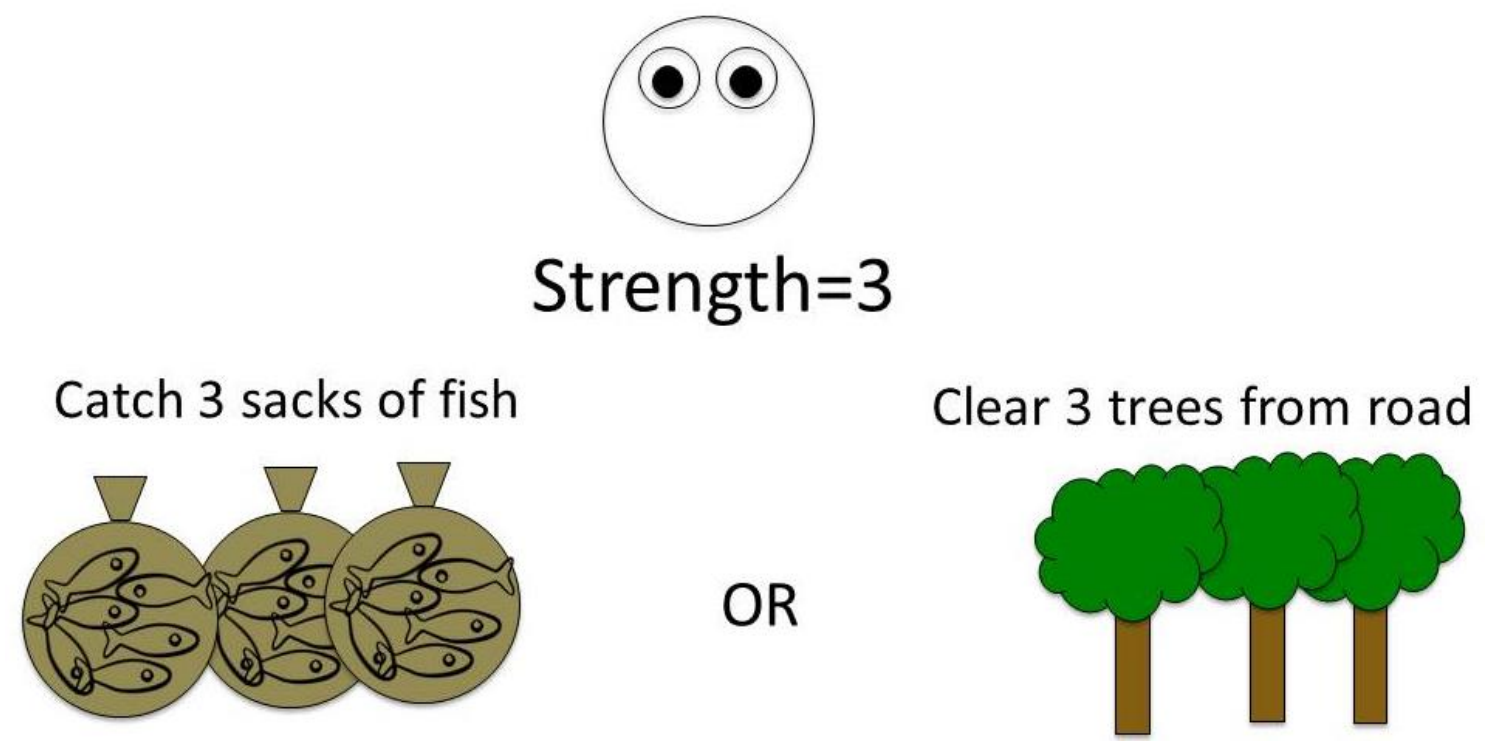
- In **hindsight**, how important was the choice of the person in this scenario?
- Requires the use of **counterfactuals** to compare the current world with ones in which some agents' choices are modified.
- Here we use the **structural model**[2], which requires determining how many changes (N) to the current scenario would be necessary to make a specific agent's actions pivotal for the outcome.
  - $Blame = \frac{1}{N+1}$

Experiment Overview

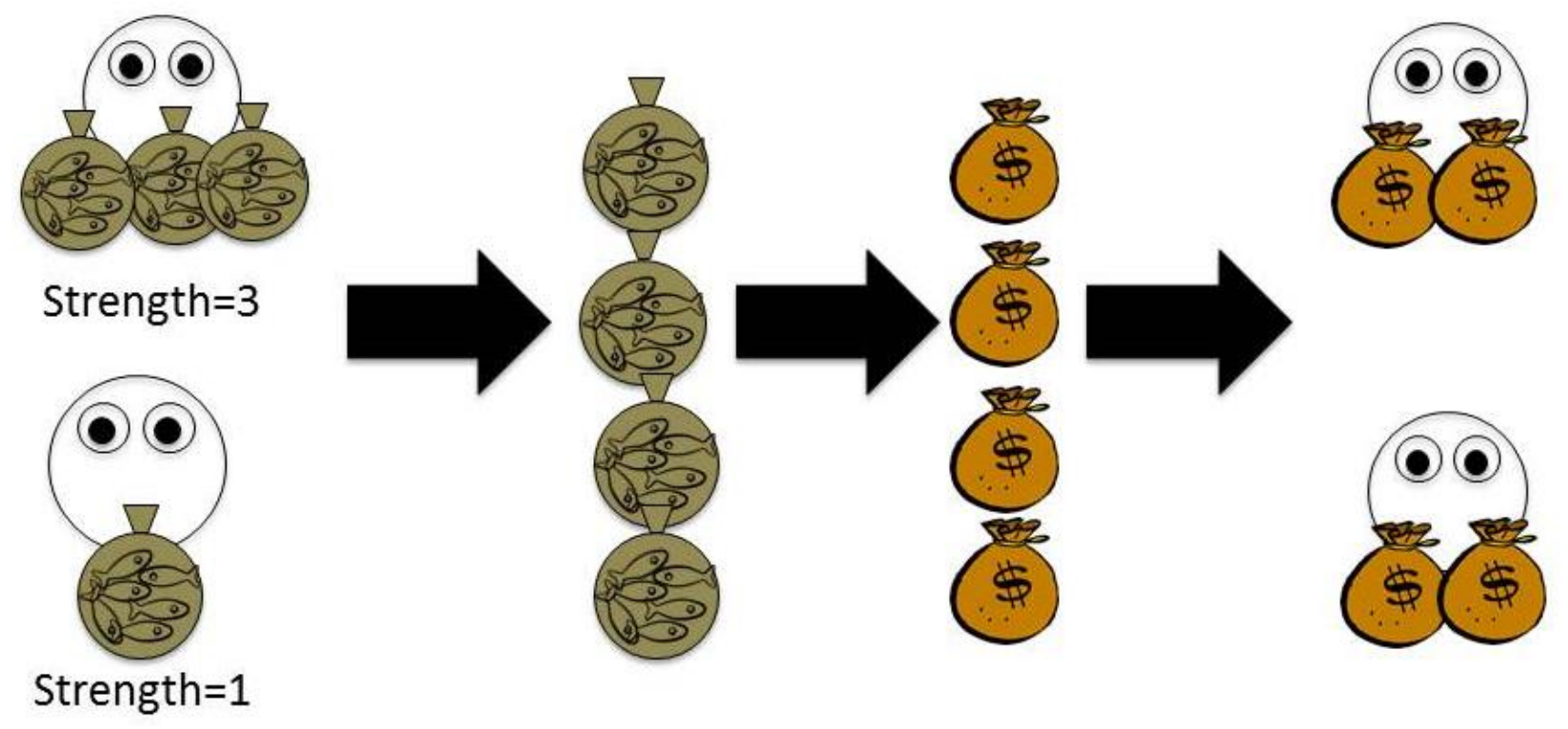
Three fishermen live in a village with a trading route often blocked by trees.



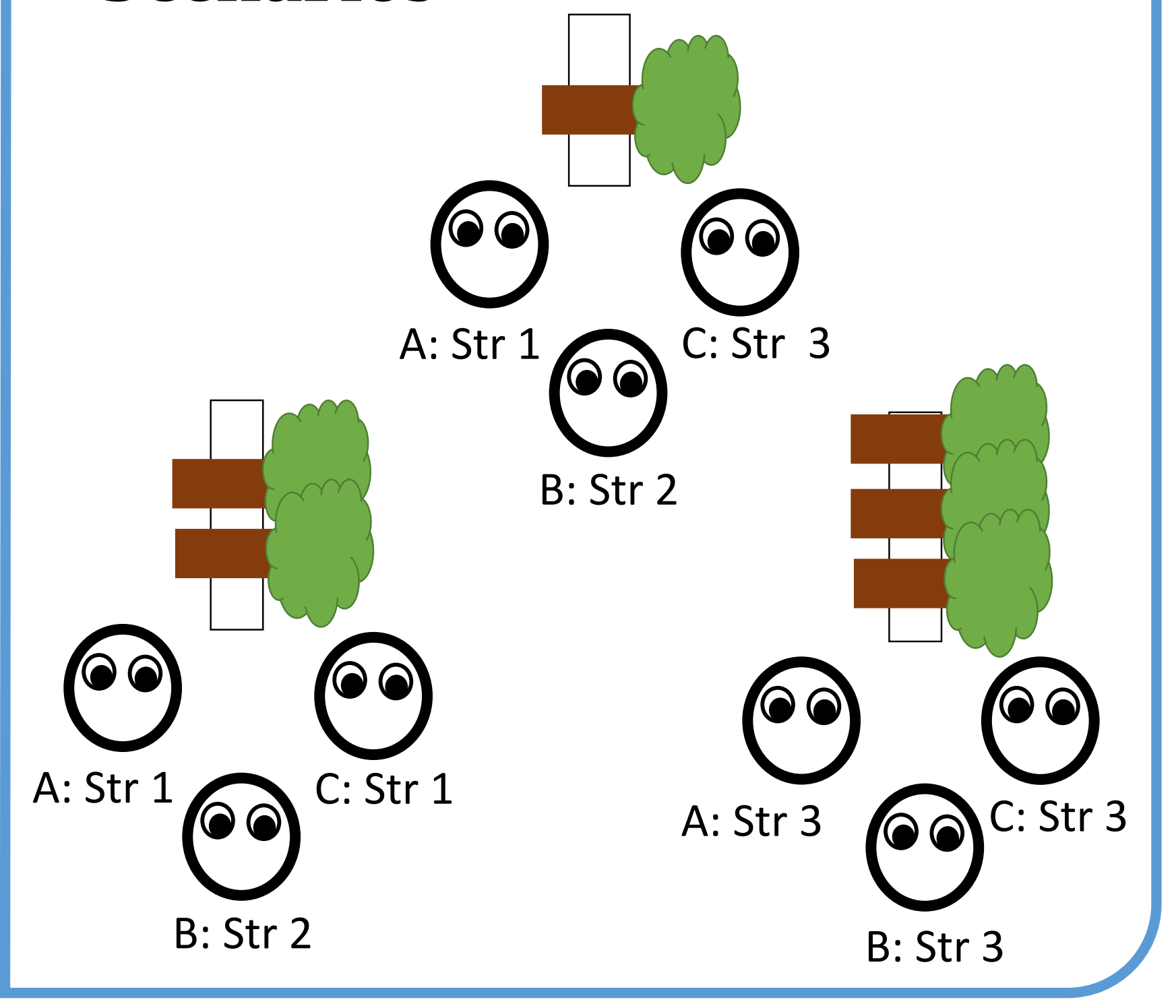
Each fisherman has a different strength, which determines how many fish he could catch and how many trees he could clear.



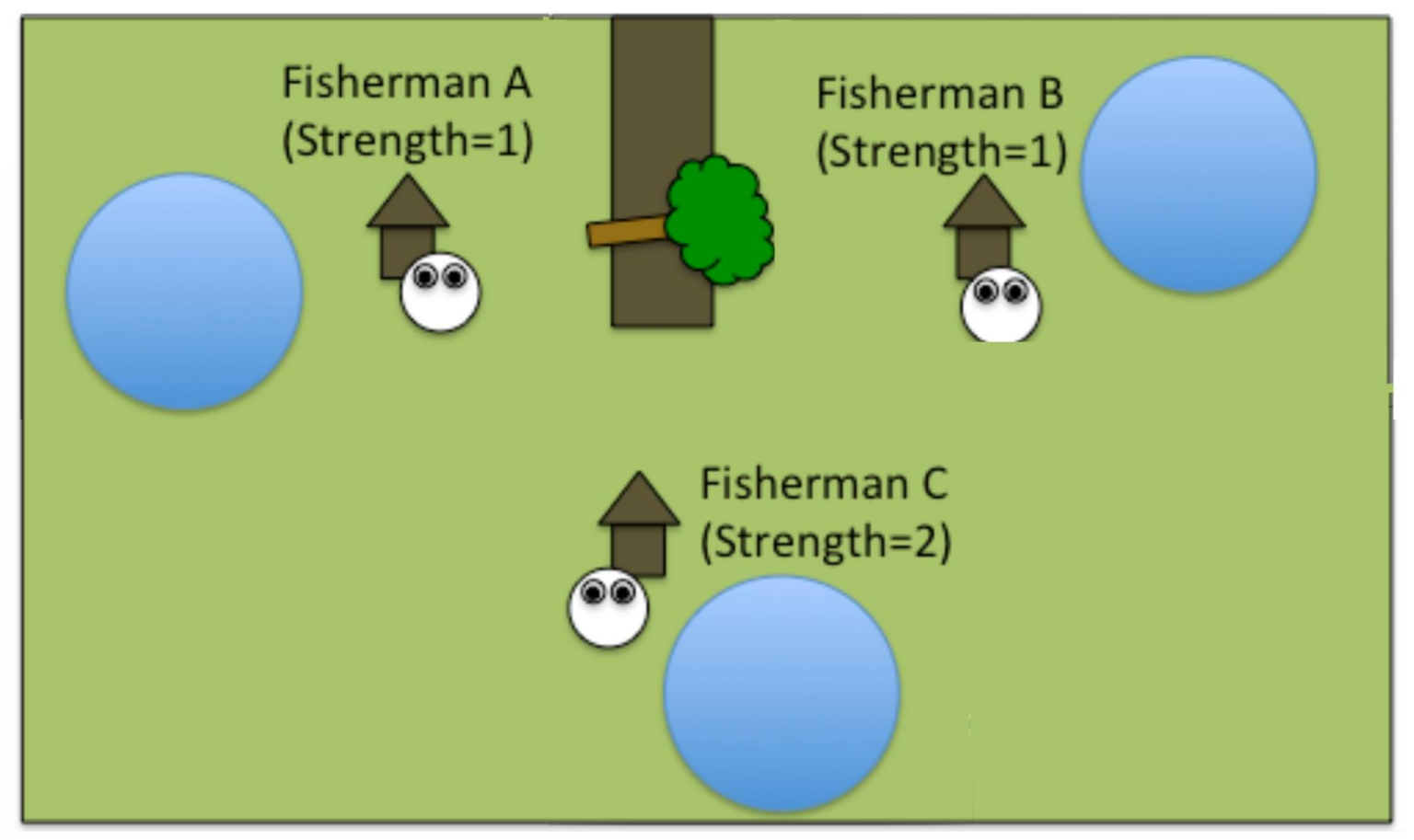
At the end of the day, if the trees have been cleared from the road, the fishermen split their earnings from the fish equally



Scenarios



Experiment 1



What should Fisherman A do?

Rationality

Recursive Rationality

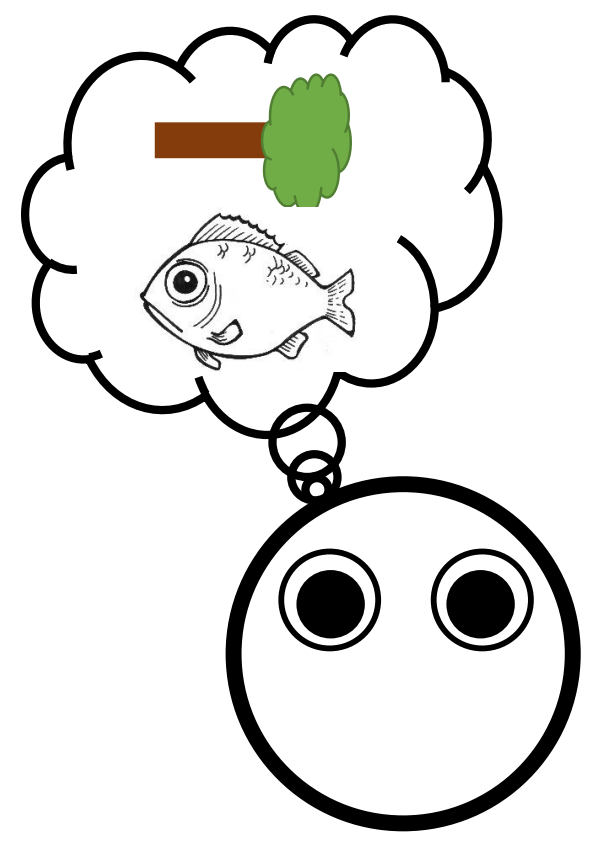
Fishermen try to predict the actions of the other fishermen, and use this knowledge to plan their own action.

- A fisherman i takes action a_i with probability $p^k(a_i)$.
- $\hat{r}_k$ is the expected reward for the action a_i of fisherman i at level k.
- R describes the rewards the fishermen could receive for every action combination.
- β is a rationality parameter.

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

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

Uniform over Best

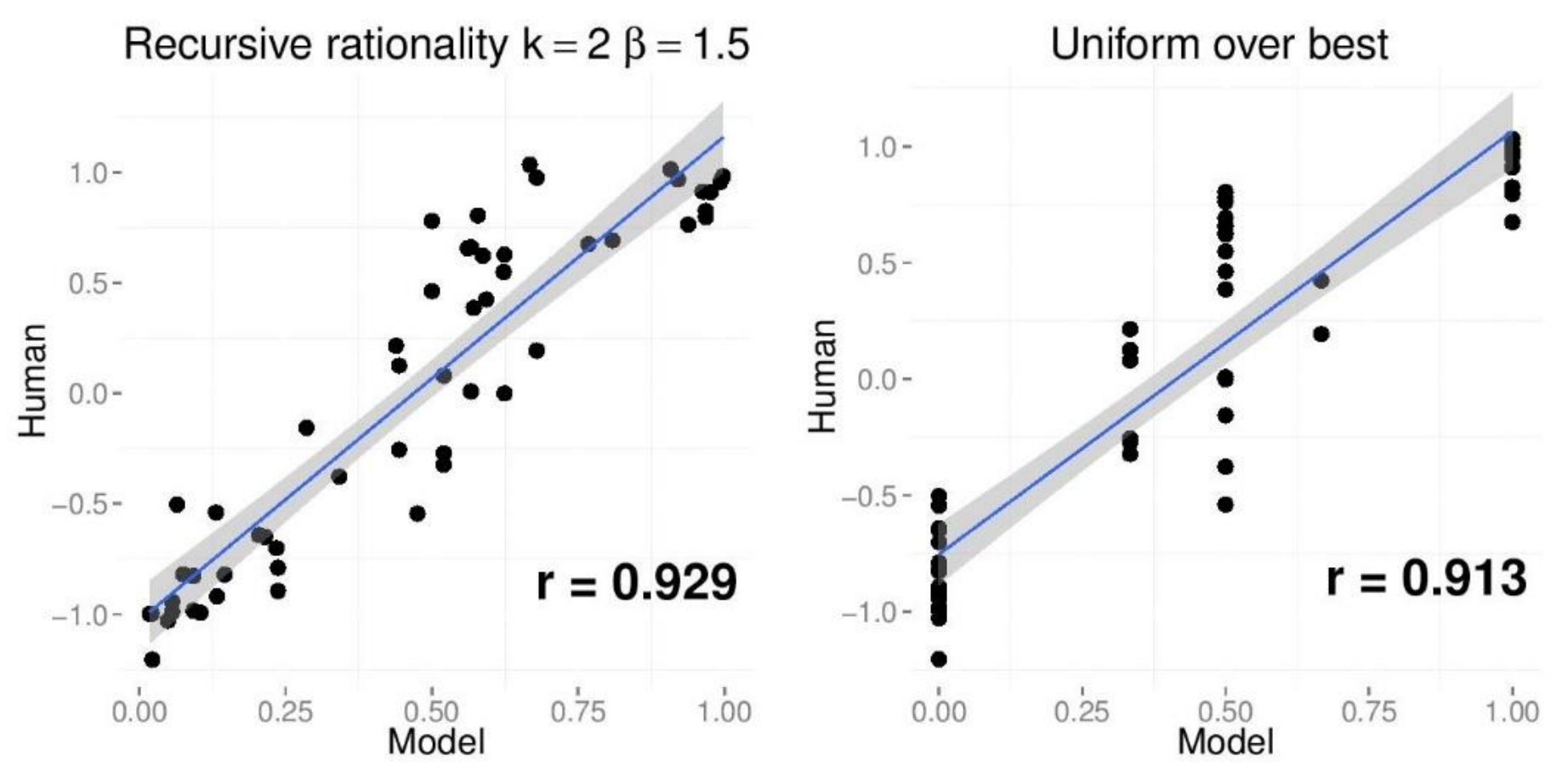


Fishermen consider all optimal scenarios and choose an action in proportion to its frequency in these worlds.

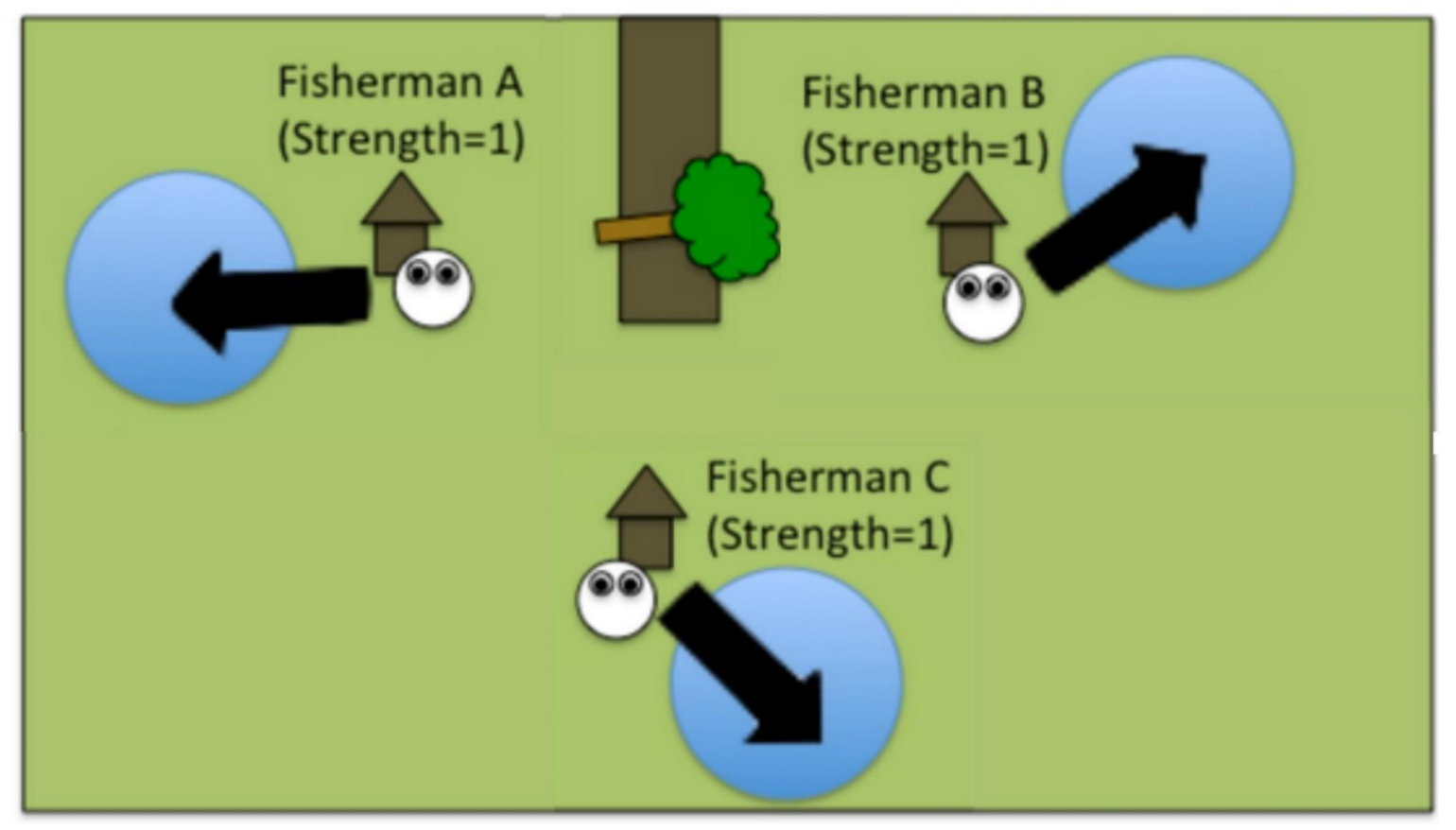
$$p(a_i) = \frac{R_{opt}(a_i)}{R_{opt}(a_i) + R_{opt}(\sim a_i)}$$

- $R_{opt}(a_i)$ is the number of situations in which a_i leads to an optimal reward.

Correlations with Data



Experiment 2



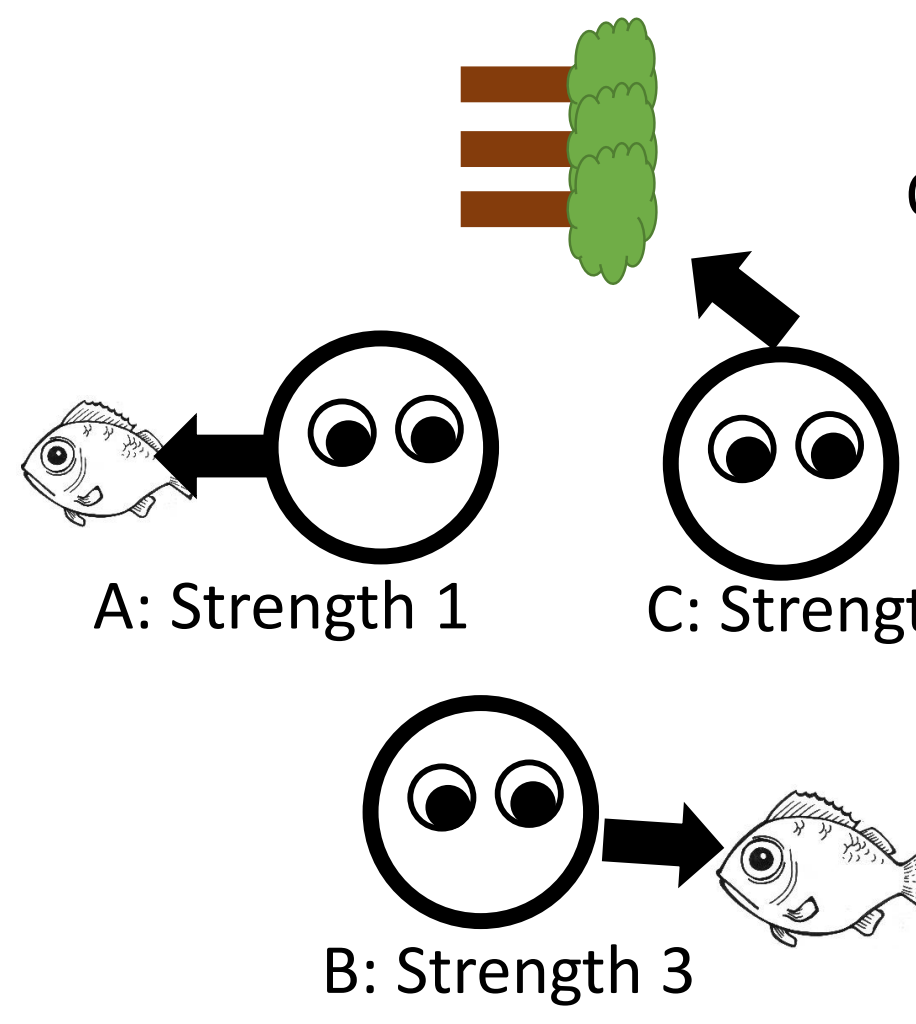
How much is each fisherman to blame for **the group's** failure to get the best possible outcome?

Pivotality

Optimal Pivotality

Determined by the number of changes needed to make a fishermen pivotal in the closest possible optimal world.

$$Pivotality_{opt} = \frac{1}{N_{opt} + 1}$$



Optimal Pivotality
Fisherman A: 1
Fisherman B: 0
Fisherman C: 0

Blame

Rationality Only

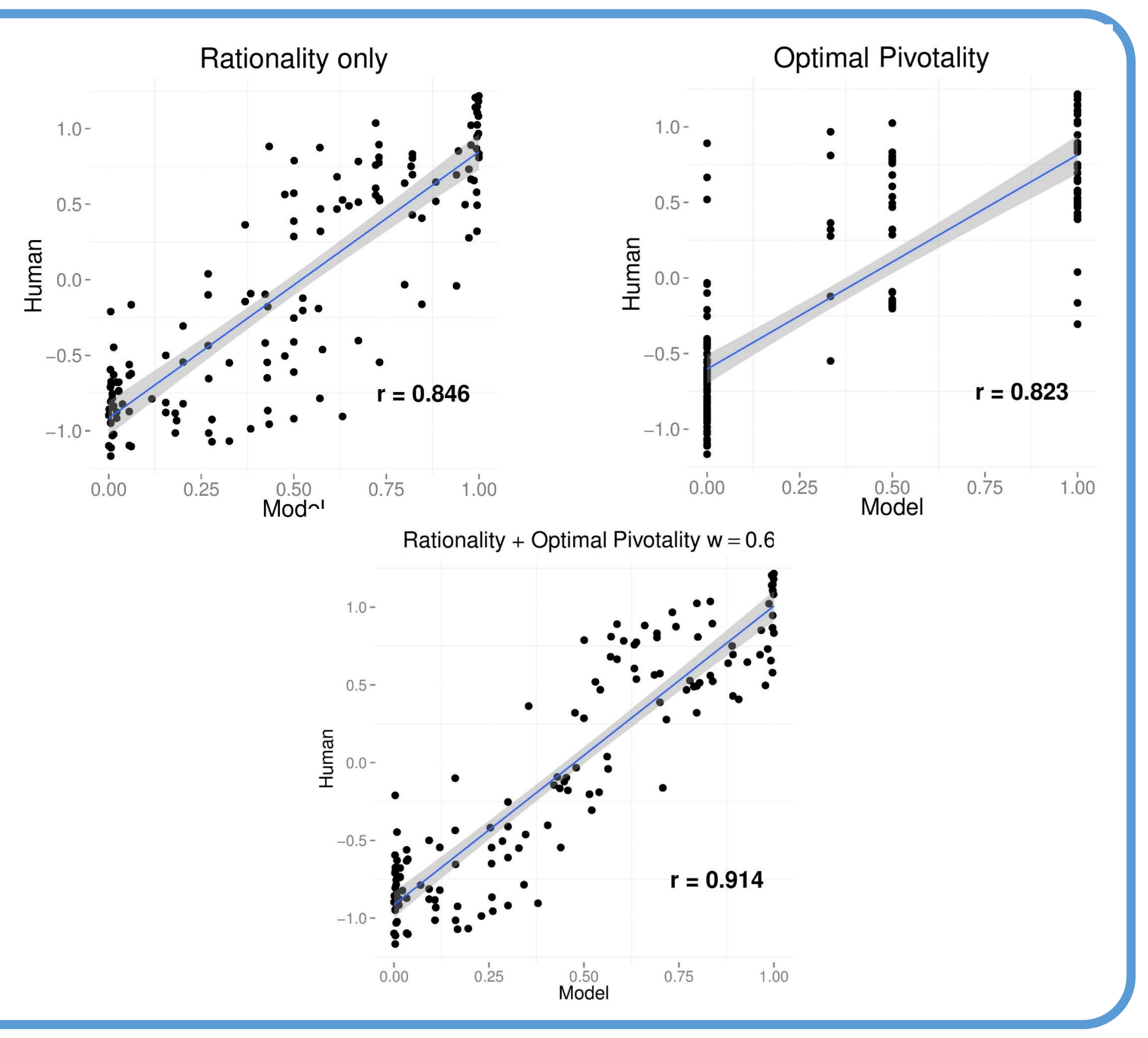
$$Blame_i = 1 - p^k(action_i^*)$$

Optimal Pivotality

$$Blame_i = \frac{1}{N_{opt} + 1}$$

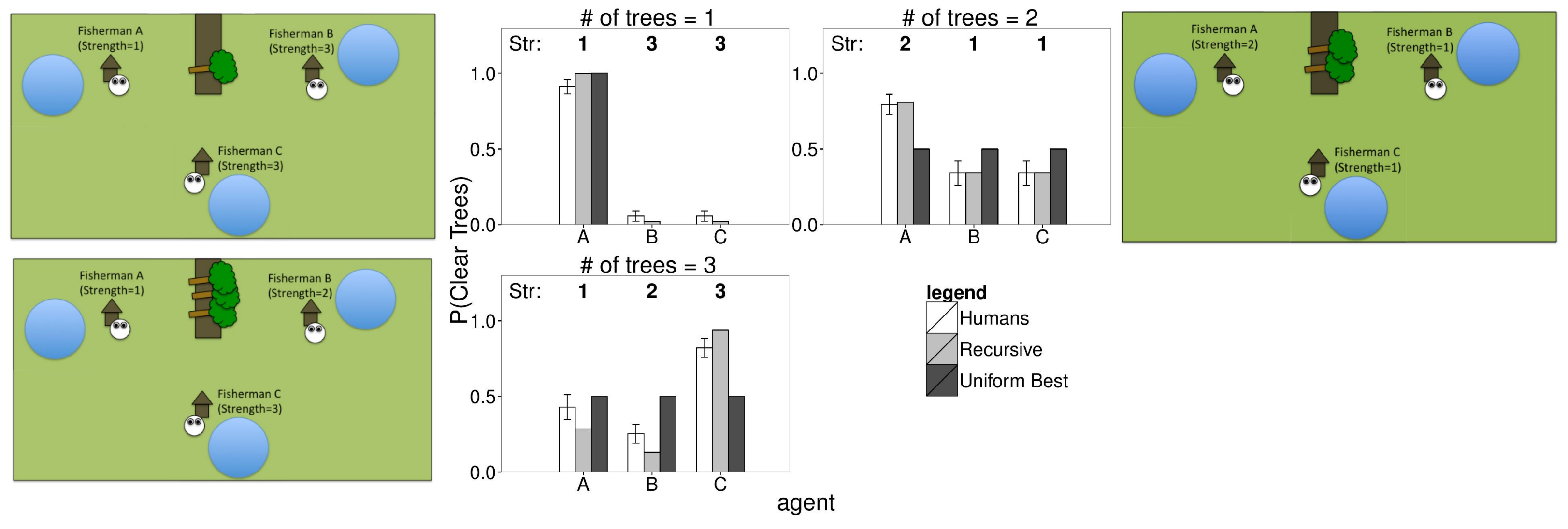
Mixture Model

$$Blame_i = w \times Rationality + (1 - w) \times Optimal\ Pivotality$$

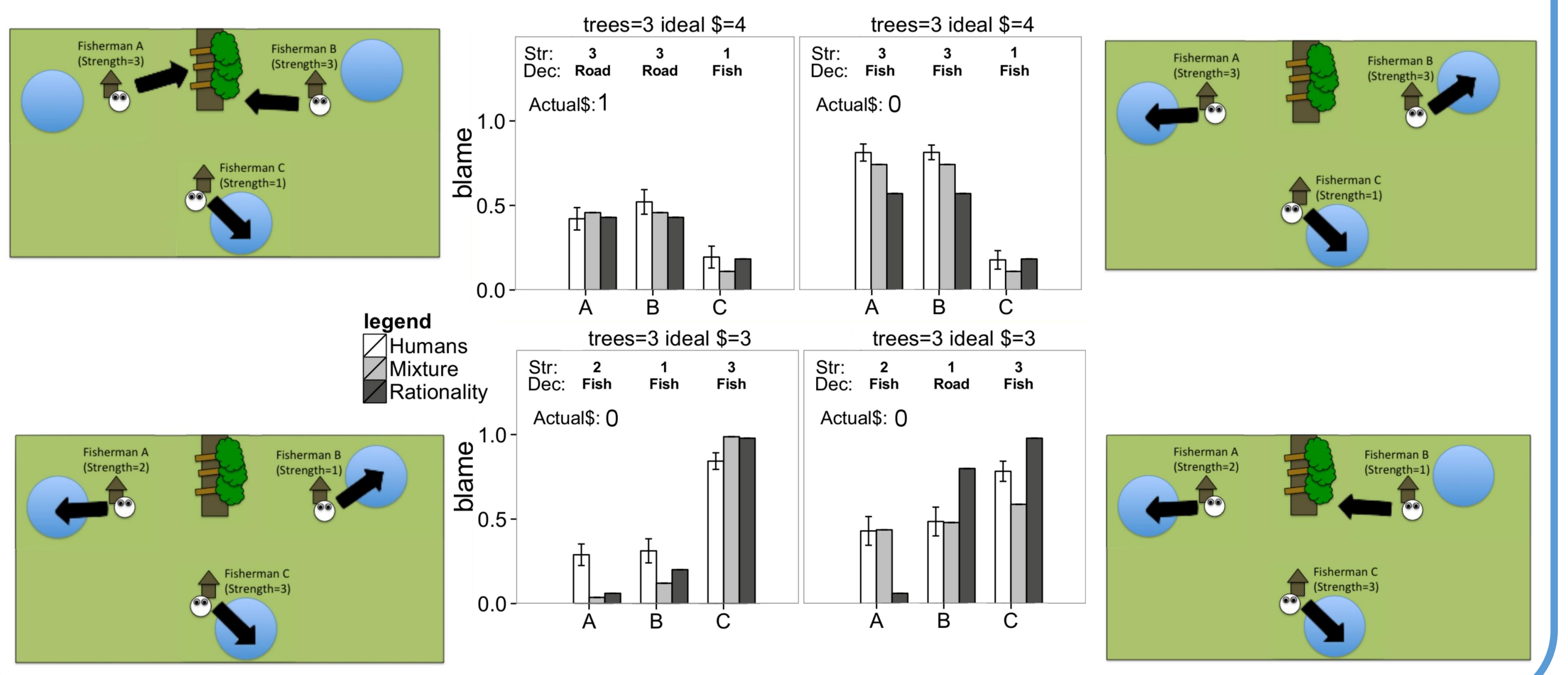


Detailed Model Comparison

Action



Blame



Discussion

- Both person-centric and action-centric measures of responsibility are important when attributing blame to individuals in a group.
- The person-centric aspect of responsibility is derived from the assumption that the other group members behave rationally (here as a depth 2 recursive reasoner).
- Future work will look at how we establish social norms through repeated interactions with the same group, and how this affects our judgments of blame and credit.

Acknowledgements

This work was funded by the Center for Brains, Minds and Machines through 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

[1] Allen, K., Jara-Ettinger, J., Gerstenberg, T., Kleiman-Weiner, M. & Tenenbaum, J. B. (2015). Go fishing! Responsibility judgments when cooperation breaks down. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

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

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