Spreading the Blame – The Allocation of Responsibility amongst Multiple Agents

Tobias Gerstenberg & David Lagnado, University College London
contact: tobias.gerstenberg@gmail.com

How would you Spread the Blame?

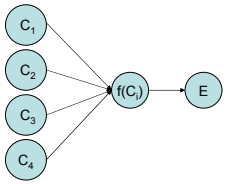
Losing football team

Bankrupt investment bank

Marksmen of a firing squad

Problem Situation

How much is an individual agent C_i responsible for a collectively brought about outcome E?



?

Is the degree of responsibility dependent on the way in which the contributions of the individual causes are combined $f(C_i)$ to determine the collective outcome?

Proposed Solutions

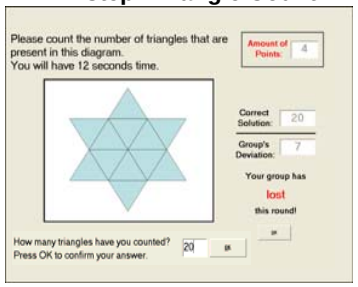
1. **The individual contribution** (Matching Model)
   Agent receives responsibility according to his individual contribution to the collective outcome.
2. **The potential of making a difference in the *actual situation*** (Simple Counterfactual Model)
   Agent receives responsibility if he had the potential of changing the collective outcome by having acted differently.
3. **The potential of making a difference in a *possible situation*** (Minimal Change Model)
   Agent receives responsibility according to the minimal number of changes N that would have been needed to alter the actual situation so that the action of the agent would have become critical to the collective outcome.
   Degree of Responsibility of C_i for E = $\frac{1}{N+1}$

Methodology

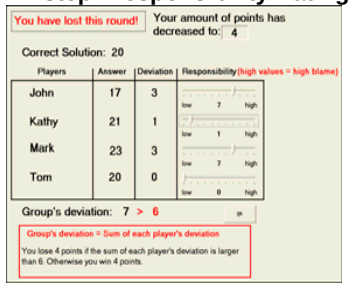
In order to evaluate the fit of the proposed solutions with people's intuitions of how responsibility should be assigned, we developed an experimental game – the Triangle Game.

In the Triangle Game, participants form a team together with three computer players. Whether the team wins or loses a particular round depends on the performance of each player in the team. The game is played for ten rounds, with each round consisting of the two following steps.

1st step: Triangle Count



2nd step: Responsibility Rating



1st step: Participants count the number of triangles in briefly presented diagrams. They then see the group's result in this round.

2nd step: Participants assign responsibility to the individual players of their group for the collective result in that round.

3 Experimental Conditions:

The only way in which the three between-subject conditions differed was in terms of the underlying rule which determined whether a round was lost or won dependent on the individual players' deviations.
The team won if the …

1. sum: *sum* of the players' deviations ≤ **6**.
2. least: deviation of the *least* accurate player ≤ **2**.
3. most: deviation of the *most* accurate player = **0**.

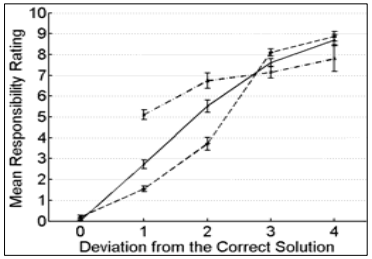
Example Situation:

Player	Deviation
John	0
Kathy	1
Mark	3
Tom	2

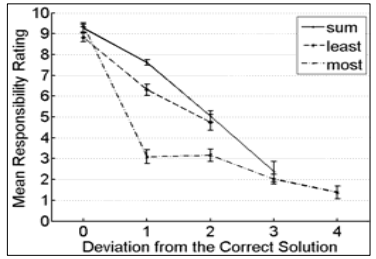
Win if
1. sum ≤ **6**
2. least ≤ **2**
3. most = **0**

Group's Deviation	Result
6	win
3	loss
0	win

Results



Losses



Wins

The different experimental conditions had a significant effect on how people assigned responsibility depending on the accuracy of an agent's answer.

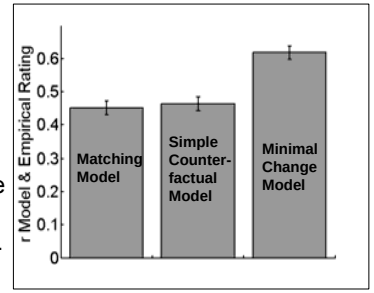
But what were the strategies that people used in order to assign responsibility? The predicted responsibility ratings of the proposed solutions were compared against the participants' empirical ratings.

Model Predictions:

1. Matching Model:
Direct match between a player's deviation and the responsibility rating.

2. Simple Counterfactual Model:
Checks counterfactual dependence between a player's answer and the team's result in the actual situation.

3. Minimal Change Model:
Checks counterfactual dependence between a player's answer and the team's result under a minimally altered situation.



Discussion

1. Participants' responsibility ratings were best predicted by the Minimal Change Model.
2. Participants' attributions of responsibility were sensitive to the way the individual contributions were combined to determine the collective outcome of the whole group.

Gerstenberg, T. & Lagnado, D. A. (under review). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*.
Chockler, H. & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22, 93-115.