The Dice are Cast:
The Role of Intended versus Actual Contributions in Responsibility Attribution

Tobias Gerstenberg (t.gerstenberg@ucl.ac.uk), **David A. Lagnado (d.lagnado@ucl.ac.uk)**
Department of Cognitive, Perceptual, and Brain Sciences
University College London, United Kingdom

Yaakov Kareev (kareev@vms.huji.ac.il)
Center for the Study of Rationality
The Hebrew University of Jerusalem, Israel

Abstract

How much are people's responsibility attributions affected by intended versus actual contributions in group contexts? A novel experimental-game paradigm dissociated intended from actual contributions: good intentions could result in bad outcomes and bad intentions in good ones. Participants acted as external judges and attributed responsibility to computer players engaging in a repeated game. On each round, three players formed a group and each chose to roll one of three dice that differed in terms of price and probability distribution. The team won if the sum exceeded a certain threshold. The results showed that both intended contribution, reflected in the choice of die, and actual contribution, reflected in the outcome of rolling the die, were determinants of participants' responsibility attributions. However, contrary to previous evidence (Cushman, Dreber, Wang, & Costa, 2009), more participants based their attributions on the intention rather than the outcome.

Keywords: responsibility; attribution; intentionality; outcome bias; experimental game.

Introduction

At the beginning of the movie "Naked Gun 2 1/2" the police officer Frank Drebin is honoured at the presidential dinner for his recent achievement of having eliminated his 1000^{th} drug dealer. In response to this, Mr Drebin admits that he had run over the last two men with his car. Luckily, it turned out that they were wanted drug dealers. Cases of "moral luck" have drawn the attention of philosophers (Williams, 1981), legal scholars (Hart, 1985), and psychologists (Mitchell & Kalb, 1981). These situations are characterized by the fact that the outcome of an action influences its moral evaluation retrospectively, even if this outcome was to a large extent beyond the control of the agent. Mr Drebin, for example, receives praise for his reckless driving only because the men he ran over happened to be drug dealers: a circumstance which was clearly beyond his control.

That people are influenced by outcome knowledge is a well established psychological finding (Baron & Hershey, 1988; Fischhoff, 1975). Fischhoff (1975) showed that people are prone to a hindsight bias: knowledge about the real outcome influences the perceived likelihood of different possible outcomes. Furthermore, people appear to be unaware of the influence that outcome knowledge exerts on their judgments and are, hence, unable to control for its effect. Baron and Hershey (1988) showed that outcome knowledge influences how people evaluate decisions made under uncertainty. Even when participants had all information relevant to the decision, including the probabilities of each possible outcome, knowledge of the actual outcome nevertheless influenced their judgments of the competence of the decision-maker. Interestingly, when asked whether they *should* take the outcome into account, most participants answered in the negative.

Differential evaluations of identical decisions or actions are also reflected in the Law's differential treatment of negligence versus negligence that leads to harm, as well as cases of attempted murder versus murder. The latter cases share the fact that the person had the intention to kill; however, only in the case of murder did the intended event come about. Recently, experimental philosophers have put forward the claim that the folk notion of intention is deeply intertwined with the (moral) evaluation of the potential outcomes (Knobe, 2003). Whether a behaviour is thought to have been performed intentionally depends crucially on the outcome of that action. An identical action is judged by more participants as intentional when its outcome is blameworthy as opposed to praiseworthy.

Psychologists have also shown that intentions play a significant role when it comes to attributions of responsibility (Lagnado & Channon, 2008) and intentionality thus constitutes an important building block of psychological frameworks of responsibility attribution (Alicke, 2000; Shaver, 1985). Shaver's (1985) theory of blame assumes a linear process starting from considerations about causality, intentionality, foreseeability and potential justifications and leading to judgments of blame or praise. In contrast, Alicke's (2000) account acknowledges the possibility of that process being reversed. The valence of the outcome can trigger spontaneous moral or emotional evaluations which influence the perception of the antecedents of the outcome. This includes judgments about the causal impact of an agent, whether the action was performed intentionally as well as whether the agent should have foreseen the outcome.

The importance of the concept of intentionality has also been recognized by economists. Variations of classic economic games, like the ultimatum game, have been employed to investigate the effects of outcome versus intention on people's perception of fairness. In the ultimatum game (Güth, Schmittberger, & Schwarze, 1982), a first player is allocated a certain amount of money. She can then decide how much of that money to give to a second player, who can either accept or reject the offer. If he refuses, both players get nothing.

Two main findings with respect to the influence of intentions on the behaviour of the second player are worth mentioning. First, if the allocation of the first player is determined by a computer and hence cannot be ascribed an intention, the rejection rates for "unfair" offers are significantly lower (Falk, Fehr, & Fischbacher, 2008). Second, the rejection rates of unequal offers strongly depend on the allocator's set of possible alternatives (McCabe, Rigdon, & Smith, 2003). The acceptability of an action is hence evaluated with respect to the choice set and an unequal offer more readily accepted if the allocator could not have been kinder. In order to accommodate these findings, economists have moved from fairness theories that only considered outcomes (Fehr & Schmidt, 1999) to theories based on intentions (Dufwenberg & Kirchsteiger, 2004) and theories incorporating both intentions and outcomes (Falk et al., 2008).

As demonstrated by the moral luck example in "Naked Gun", there is another factor beyond intentions and outcomes that is relevant when it comes to considerations about fairness or attributions of responsibility: the control an agent has over the outcomes he brings about. Our environment is fundamentally noisy and, most of the time, we only have partial control over the effects of our actions. While it is true that the valence of intention and outcome are correlated in everyday life, this relationship is imperfect. Good intentions can sometimes lead to bad outcomes and bad intentions to good ones. For example, a careful driver might cause the death of a careless child. In order to understand the complex relationship between intentions, outcomes and control it is necessary to create experimental situations in which these factors can be dissociated.

In a recent study, Cushman et al. (2009) investigated the effects of intention versus outcome on perceived fairness in a two-player, allocater-responder game. Similar to the ultimatum game, the allocator proposed how a pot of \$10 should be shared. Allocations were either stingy (player 1: \$10, player 2: \$0), fair (\$5, \$5) or generous (\$0, \$10). The responder could punish or reward the allocation of player 1 by subtracting or adding up to \$9 to her account. Importantly, in one condition of the experiment, the allocator only had partial control over the outcome. She had to choose which one of three possible dice she wanted to roll. These dice differed in terms of the probability with which they led to stingy, fair or generous outcomes. Following a strategy format, responders had to indicate for each of the 9 possible combinations (e.g. generous die, stingy outcome) how much money they wanted to add or subtract from the allocator. The results revealed that participants were much more influenced by actual outcomes than by intentions. Responders tended to subtract money for selfish outcomes for all three dice, whereas they added money for fair and generous outcomes. The choice of die exerted only a small effect on this general pattern. Surprisingly, the results of a condition in which the allocator had perfect control were virtually identical. Hence, the study provides further support for the finding that people can be so sensitive to outcomes that they sometimes disregard the underlying intention that lead to that outcome. However, Cushman et al. (2009) admit that methodological limitations might have contributed to their findings. Importantly, since the responder is part of the game it is the outcome and not the intention that is the most relevant to him. In order to validate their findings, it is important to investigate how an independent judge would have decided.

The current experiment addressed this limitation. We created a setting in which an external observer evaluated the behaviour of agents participating in an experimental game. The following scenario helps to exemplify the main components of our experiment: Sarah is running for the position of student representative. Three friends are helping her campaign by distributing flyers. Tom puts in a lot of effort and distributes 100 flyers. John puts slightly less effort into the campaign and only distributes 50. Finally, Alex thinks that Tom's and John's contributions are probably already enough to win the campaign and he only distributes 30 flyers. As it turns out, 20 of Tom's, 20 of John's and 25 of the people who received their flyer from Alex voted for Sarah. As a result, Sarah won the election. Assuming that Sarah knows about both the number of distributed flyers and the votes she received, how much is she going to praise each of her three friends for their contribution to her win?

Two aspects of the outlined scenario are important with respect to the current study. First, it shows how intention and outcome can sometimes mismatch in situations over which agents exercise only partial control. Despite Tom's good intention and effort he contributed no more to the collective outcome than John and even less than Alex. Second, the scenario entails a component that is characteristic of social dilemmas (see e.g. Hardin, 1968). Each individual agent has to weigh the cost of the effortful process of distributing flyers with the potential gain of an election won. Alex's thought process indicates each person's motivation to free-ride on the effort of the others. Assuming the spoils of a victory are equally shared, the person who put in the least effort will have the highest net benefit.

The current study investigated the effects of intended and actual contributions on responsibility attribution in a group context in which agents had only partial control over their contributions. How well can intended contributions, actual contributions, or their combination explain participants' responsibility attributions?

Experiment

The aim of the experiment was to generate a situation in which intended versus actual contributions could dissociate. Participants took the perspective of an external observer and judged the behaviour of computer players engaging in an experimental game (see Figure 1). Hence, participants did not actively engage in the game themselves. In each round of the game, three computer players were randomly selected to form a group. Each player chose one of three available dice to roll. The dice differed in terms of their underlying probabil-

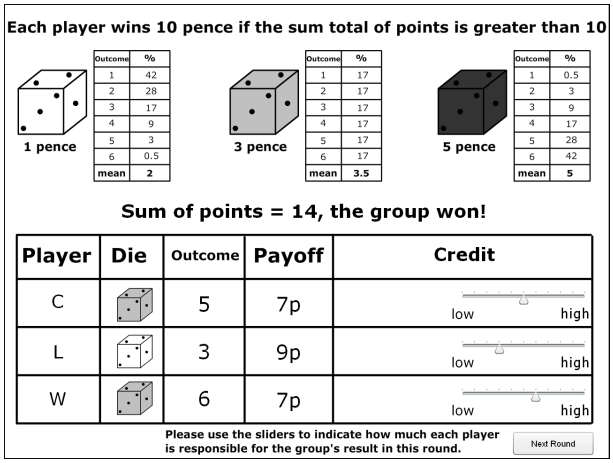

Figure 1: Screenshot of the game depicting a won round.

ity distributions (see top part of Figure 1). The white die had a higher probability of smaller values, the grey die was fair, and the black die was skewed towards higher values (in the experiment, the colours were bronze, silver and gold). The group of players won a round if the sum of their outcomes was greater than 10. If the group won a round, 30 pence were equally distributed between the players. If the group lost, no money was distributed. Importantly, the players had to pay different amounts for the dice before they rolled them. The white die cost 1 pence, the grey die 3 pence, and the black 5 pence. Each individual player's payoff was a function of the group's result, that is, whether they won or lost, and the money he had to pay for the die of his choice. The task of the participants as independent judges was to indicate how much they thought each player was responsible for the group's result in each round.

The computer players chose each of the dice with an equal probability. The chosen payoff function created a social dilemma. The overall probability of winning was 50%. The probabilities of winning given that a player had chosen the white, grey or black die were 33%, 50% and 68%, respectively. This led to an expected payoff of 2.3 pence per round for the white die ($33\% \times 9 - 67\% \times 1 = 2.3$). The expected payoffs for the grey and black die were 2 and 1.8 pence. Hence, there was an incentive for each player individually to choose the white die. However, if all of the players chose that die, the probability of the team winning was only 2%, and the expected payoff -0.8 pence.

Figure 2 shows the underlying structure of the experiment. The choice of die reflected the *intended* contributions of the players while the team's result was a function of the *actual* contributions. We predicted a main effect of intention: the same outcome of roll will elicit different responsibility attributions dependent on the choice of die. We also predicted a main effect of outcome: responsibility attributions for a given die will vary with the outcome of rolling this die. Finally, previous research suggested that outcomes will affect participants' responsibility ratings more strongly than intentions (Cushman et al., 2009).

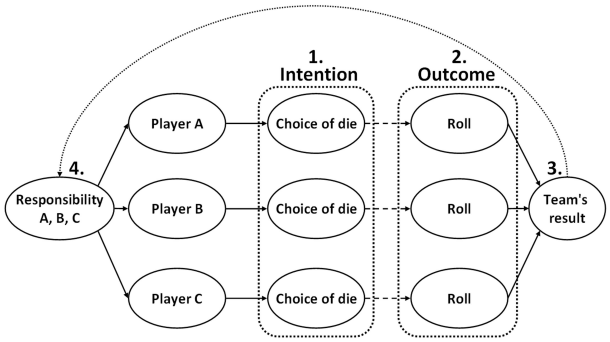

Figure 2: Underlying structure of the paradigm. Numbers 1. - 4. indicate the different components of each round.

Method

Participants and Materials 80 participants from the UCL subject pool participated for the chance of winning one of six amazon vouchers worth £150 in total. 55 participants were female and the mean age was 23.2 (5.94). With the second part of the experiment added at a later stage (see Procedure), 35 participants performed only the first part of the experiment, whereas the remaining 45 participants performed both. The study was conducted online and programmed with Adobe Flash.

Procedure Participants were informed that the experiment would take 20 minutes and that their task was to evaluate the behaviour of players engaged in an experimental game by attributing credit for wins and blame for losses. Participants read a description of the three dice which made it clear that they differed in terms of both probability distribution and price. A practice round served to familiarize participants with the structure of the game. After the practice round, participants had to answer questions to ensure that they had understood the rules of the game correctly. The game was then played for 20 rounds.

On each round, participants saw a table that showed for each player which die she had chosen, the outcome of having rolled that die and the amount won or lost in that round. In Figure 1, Player *C* chose the grey die and rolled a 5. Her payoff was 7p since she paid 3p for rolling the grey die and each player received 10p for winning this round. Players were indicated by capital letters which changed in each round. This was done to prevent participants from forming an overall impression about individual players. The header above the table showed the sum of points and changed its colour from green to red according to whether the round was won or lost. For each player, participants attributed blame for losses or credit for wins, by moving a slider ranging over a scale from 0-10. Its endpoints were labelled 'low' and 'high'. In line with the result of the round (loss/win), the label (blame/credit), color (red/green) and position of sliders (middle to left/ middle to right) of the last column changed.

45 of the 80 participants also completed a second stage of the experiment. Those participants were informed after the 20^{th} round that they would see 14 novel situations that could

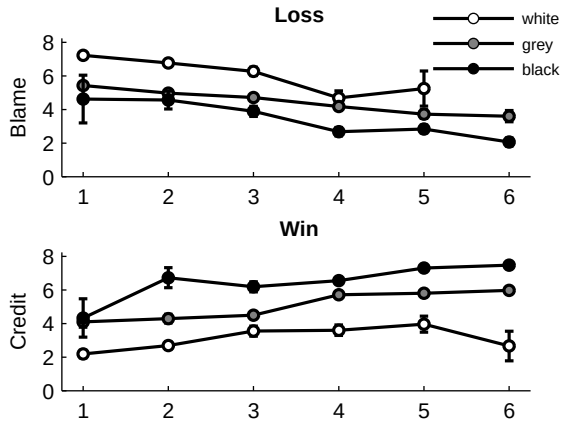

Figure 3: Mean responsibility ratings for each combination of die and outcome for both losses and wins. Lines represent the different dice and values on the x-axes indicate the outcome of rolling each die.

have occurred in the game and which were of special interest to the researchers. As explained below, the test cases were chosen so as to enable a fine assessment of the weight assigned to intentions and outcomes. The order of these test cases was randomized. Finally, participants were asked to indicate in a textbox whether they had focused on the choice of die, the outcome, or both.

Results

Mean Responsibility Ratings

In order to evaluate the effects of choice of die and outcome of roll for the first stage of the experiment, we ran separate 3 (Die) x 6 (Roll) ANOVAs for both wins and losses. Figure 3 shows the mean responsibility attributions as a function of choice of die and outcome of roll. For wins, there was a significant main effect of Die $F(2, 2472) = 87.10$, $p < .001$, $\eta^2 = .066$ and of Roll $F(5, 2472) = 9.53$, $p < .001$, $\eta^2 = .019$, as well as an interaction effect $F(10, 2472) = 1.91$, $p < .05$, $\eta^2 = .008$. For losses, there was a significant main effect of Die $F(2, 2327) = 31.62$, $p < .001$, $\eta^2 = .027$ and of Roll $F(5, 2327) = 15.31$, $p < .001$, $\eta^2 = .032$, but no interaction effect ($p > .05$).

To qualify these results, we ran linear contrasts on Die and Roll for both wins and losses. For wins, there was a significant positive linear trend of Die as well as for Roll. For losses, there was a significant negative linear trend of both Die and Roll (all p's $< .001$).

These analyses show that overall, both the choice of die and the outcome of its rolling influenced participants' responsibility ratings. However, the results cannot reveal how individual participants weighted these two factors. To find out, we conducted individual regression analyses, and report them below.

Regression Analysis

First, we ran the following three separate regression analyses based on the overall data (80 participants x 20 rounds x 3 ratings data points):

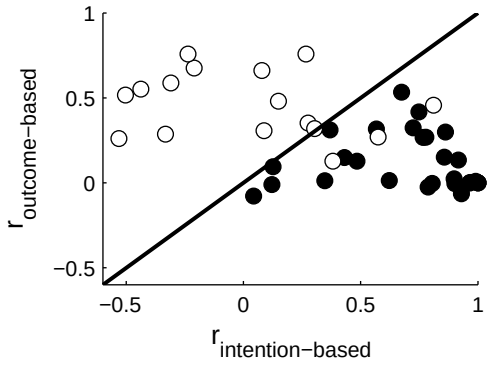

Figure 4: Scatterplot of correlations with outcome-based model and intention-based model. Black circles indicate participants classified as intention-based (N = 29), white circles indicate participants classified as outcome-based (N = 16).

$$\text{intention-based model: } responsibility = \beta_0 + \beta_{1\ die} \quad (1)$$

$$\text{outcome-based model: } responsibility = \beta_0 + \beta_{1\ roll} \quad (2)$$

$$\text{mixture model: } responsibility = \beta_0 + \beta_{1\ die} + \beta_{2\ roll} \quad (3)$$

All three regression models accounted for a significant amount of the variance in the data (see Table 1).

Evaluation of Test Cases

To break the results down even further, we ran the regression models for each individual participant. Based on the magnitude of the correlation with the intention-based regression model versus the outcome-based regression model, we grouped the 45 participants who completed the second stage of the experiment in two groups. We used this grouping to predict how participants would attribute responsibility for the chosen test cases (described below). Figure 4 shows how well the behaviour of the classified participants was predicted for the test cases.

The test cases were constructed to enable a fine analysis of the relative weights assigned by participants to intentions versus outcomes. It should be noted that in the first 20 rounds of the experiment the choice of die and outcome of roll were highly correlated due to the chosen probability distributions ($r = .68$, $p < .001$). In contrast, for the test cases the choice of die and outcome of roll were uncorrelated ($r = 0$). This shows that these test cases indeed created situations that could

Table 1: Results of overall regression analyses.

Model	R^2	F	$p <$	β	t	$p <$
intention-based	.268	1757.70	.001	0.518^a	41.93	.001
outcome-based	.219	1346.33	.001	0.468^b	36.69	.001
mixture	.303	1042.31	.001	0.370^a	24.02	.001
				0.238^b	15.48	.001

$a = \beta_{die}$, $b = \beta_{roll}$

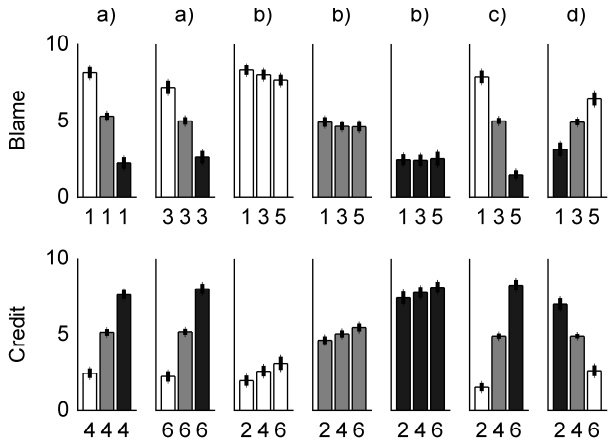

Figure 5: Mean responsibility ratings of *intention-based* participants for 14 test cases. The top row shows losses and the bottom row wins. The values on the x-axes indicate the outcome. The colours of the bars indicate the dice.

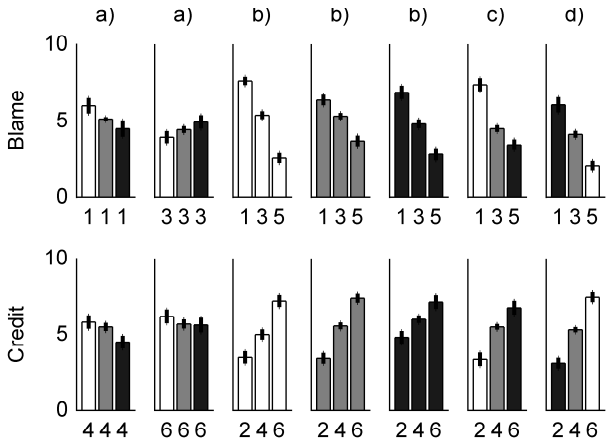

Figure 6: Mean responsibility ratings of *outcome-based* participants.

be used to distinguish between intention-based and outcome-based participants.

Figure 5 shows how the 29 intention-based participants attributed responsibility in the test cases. Figure 6 shows the responsibility attributions for the 16 participants who had been classified as outcome-based. The test cases can be categorized into 4 groups: a) different dice, same roll; b) same dice, different rolls; c) congruent; d) incongruent. 'Congruent' means that the quality of die and outcome of roll corresponded (i.e. the expensive die led to a high and the cheap die to a low outcome); 'incongruent' means that the quality of die and the outcome of roll mismatched.

Inspection of the graphs validates the original partition. First, in the congruent test cases ('c') - which serve as a manipulation check - both groups show the same pattern of attributions, with that for the intention group being steeper than that for the outcome group. For the intention test cases ('a'), the differences in attributions are large for participants in the intention group and small for participants in the outcome group. An opposite pattern of attributions is evident with the outcome test cases ('b'): there the intention group exhibits small differences and the outcome group exhibits large differences. Finally, and most interesting, the pattern of attributions reverses in the incongruent cases ('d'). Despite the fact that in these situations the expensive die led to the lowest outcome, the intention-based participants attribute the least blame to this player for the loss (Figure 5, top) and the most credit for the win (bottom). In contrast, the attributions of the outcome-based participants for these cases closely follows the number rolled, independent of the choice of die (Figure 6).

Discussion

The current study investigated the influence of intended versus actual contribution on the attribution of responsibility in a group context. We found that both intention and outcome exerted a significant influence on participants' attributions. Furthermore, we provided evidence that individuals differ in the extent to which they base their attributions on intentions or outcomes.

Our experimental procedure allowed us to dissociate intentions from outcomes and created a situation in which participants played the role of an external judge. We found that the majority of participants were better explained as having focused on intended rather than actual contributions. Methodologically, the current experiment shows that it is important to analyse the data on the level of individual participants. While on an aggregate level, it seems that participants are weighting both choice of die and outcome of roll to determine their responsibility attribution (see Figure 3), more careful analyses reveal that most participants actually tend to either focus on the intention or the outcome (see Figure 4).

At this point, we can only speculate about the factors driving these interindividual differences. Different interpretations of the notion of responsibility could have influenced participants' behaviour. Outcome-based participants might have endorsed a *causal* conception of responsibility. Accordingly, players that rolled high numbers were credited higher since their contributions caused the win. Intention-based participants, on the other hand, might have used a *moral* conception of responsibility. Hence, players were judged for their choice of die which reflected their underlying attitude towards the team. Alternatively, the results could reflect interindividual differences in the ability or motivation to mentalize. We would assume that people who find it hard to take another person's perspective are more likely to focus an the actual outcome rather than the underlying intention. We are planning to use a simplified version of the employed paradigm to test this hypothesis on a patient group. Finally, outcome-based participants' ratings might have been influenced by beliefs about the gambling-competence of players. On this view, rolling a high number with the cheap die reflects a special ability deserving credit. Some of the participants' written comments confirm the influence of such arguably non-normative considerations.

Why did we find a relatively stronger effect of intentions when previous studies postulated the existence of an outcome bias (Cushman et al., 2009)? Several differences between studies that draw their conclusions from economic games, such as the ultimatum game, and our study could potentially explain these divergent results. First of all, most of the studies in the economic literature were interested in exploring perceived fairness and not directly in responsibility attributions. Although we presume that these notions are tightly linked, it might be that considerations about fairness and responsibility can lead to different results. Second, the participants in those studies directly experienced the outcomes, while inferring the intentions of the other player was not incentivised. In our study, in contrast, participants acted as independent external judges. It is, hence, less likely that their attention was biased towards outcomes. In a future study, we aim to explore how the patterns of attribution change when participants actively take part in the game.

An important feature of the employed experimental paradigm is its potential to explore different combination functions between the individuals in the group. Gerstenberg and Lagnado (2010) have shown that the way in which individual contributions are translated into group outcomes significantly influences people's responsibility attributions. Accordingly, an identical individual contribution can lead to very different responsibility attributions as a function of the group context. While the current experiment used an additive combination function for the contributing players, we will investigate in future experiments how attributions change when the rule of the game reflects a minimum function (i.e., the group wins if no player rolls a 1) or a maximum function (i.e., the group wins if at least one player rolls a 6). Are participants more likely to focus on the actual rather than the intended contribution when the combination function is non-compensatory?

Finally, our paradigm can be used to explore how uncertainty affects responsibility attributions. In everyday life, we do not have direct access to other people's intentions. Rather, we try to infer the intention from a person's behaviour. Our paradigm allows us to model this situation. Instead of revealing all the information to the participant, we will only show the outcomes of rolling the dice but not which dice the players have chosen. We can then compare participant's ability to infer the underlying intentions from observed outcomes with an ideal Bayesian learner and evaluate in how far their attributions can be explained by their knowledge about the players.

In conclusion, the current study explored the influences of intentions versus outcomes on responsibility attributions in a group context. We found that a majority of our participants focussed on the intention rather than on the outcome. We introduced a novel experimental paradigm which is flexible enough to lend itself to the investigation of future questions that will help to disentangle the complex relationship between control, intention, outcome and responsibility attributions.

Acknowledgments

TG is the beneficiary of a doctoral grant from the AXA Research Fund. DL & TG were supported by the ESRC Centre for Economic Learning and Social Evolution; YK by the Israel Science Foundation 539/07.

References

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

Baron, J., & Hershey, J. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*, 569–579.

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a trembling hand game. *Plos One*, *5*, 1–7.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*, 268–298.

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness: Intentions matter. *Games and Economic Behavior*, *62*, 287–303.

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*, 817–868.

Fischhoff, B. (1975). Hindsight $\neq$ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal ol Experimental Psychology: Human Perception and Performance*, *1*, 288–299.

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

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*, 367–388.

Hardin, G. (1968). The tragedy of the commons. *Science*, *162*, 1243–1248.

Hart, H. (1985). *Punishment and Responsibility*. Wadsworth Publ. Co.

Knobe, J. (2003). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*, 309–324.

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

McCabe, K., Rigdon, M., & Smith, V. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization*, *52*, 267–275.

Mitchell, T., & Kalb, L. (1981). Effects of outcome knowledge and outcome valence on supervisors' evaluations. *Journal of Applied Psychology*, *66*, 604–612.

Shaver, K. (1985). *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. Springer-Verlag New York.

Williams, B. (1981). *Moral luck*. Cambridge University Press.